I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 18174

What Did We Learn from the North American Income Maintenance Experiments? New Data and Evidence on Household Behavior and Labor Supply

Chris Riddell
W. Craig Riddell

OCTOBER 2025

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 18174

What Did We Learn from the North American Income Maintenance Experiments? New Data and Evidence on Household Behavior and Labor Supply

Chris Riddell
University of Waterloo

W. Craig Riddell
University of British Columbia and IZA

OCTOBER 2025

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

What Did We Learn from the North American Income Maintenance Experiments? New Data and Evidence on Household Behavior and Labor Supply*

We re-assess the consequences of a NIT for two-parent families, utilizing hitherto untapped data. The Gary and Seattle experiments fail balancing tests. In New Jersey, Denver and Manitoba we estimate far greater labor supply responses than the current consensus, with remarkable consistency in point estimates and statistical significance across experiments, genders and countries. On the other hand, using newly collected data from archival records, we estimate substantial increases in happiness, marital satisfaction, household production, and social activities in Manitoba. We also reject the contentious finding that the NIT increased marital separations in Seattle-Denver, which is driven solely by Seattle.

JEL Classification: C93, I38, J12, J22

Keywords: basic income, Negative Income Tax, income support, labour supply, marital satisfaction, household well-being

Corresponding author:
W. Craig Riddell
Vancouver School of Economics
Faculty of Arts
University of British Columbia
6000 Iona Drive
Vancouver, BC, V6T 1L4
Canada
E-mail: craig.riddell@ubc.ca

* Financial support from SSHRC is acknowledged. We are grateful to the staff at the Archives of Canada (Winnipeg Office) for their patience and assistance with the data and, in particular, thank David Cuthbert. The newly collected raw data and replication package is forthcoming at chrisriddell.ca. Previous versions of this paper were presented at the annual meetings of the Society of Labor Economists, European Association of Labour Economics, European Economics Association, Society of Household Economics, CreAM/RFB Workshop and the International Applied Econometrics Association. We thank participants for their comments.

1. Introduction

In the 1960s and early 1970s concern about the extent of low incomes among the 'working poor' led the US government to declare 'war on poverty' and to search for policy options. An approach supported by economists on both the left (e.g. James Tobin) and right (e.g. Milton Friedman) of the profession's political spectrum was a Negative Income Tax (NIT) with a basic income guarantee and a tax-back rate for earned income above the guarantee. Robert Solow (1987, p.220) summarizes the view at that time: "There was a feeling that we were at last in a position to eliminate poverty, that it was the right thing to do, and that the direct way to do it was to transfer income to people who would otherwise be very poor. This was combined with a feeling that the existing hodgepodge of categorical transfer programs involved the bureaucracy deeply, meanly, and inefficiently in running the lives of participants." To deal with concerns that would be raised about the potential impacts on work activity of low-income families, a randomized field experiment was proposed to provide evidence on the magnitudes of these impacts, and ultimately adopted by the (then) Office of Economic Opportunity in the US government's Department of Education, Health and Welfare (Levine, 1975). Initiated in 1968, the New Jersey Graduated Work Incentive Experiment was the first of four US Income Maintenance Experiments (IMEs) carried out during this period. North of the border, similar policy concerns resulted in the Canadian federal and Manitoba provincial governments initiating the Manitoba Basic Income Maintenance Experiment (know as 'Mincome').

These five ambitious studies were the first large-scale field experiments with a randomized design carried out in economics. The results were widely publicized and debated – in his detailed survey Widerquist (2005) cites more than 200 published papers and books, and notes that there are many more publications in the form of policy papers and reviews. The concept of an NIT and evidence on its labor supply impacts became standard fare in textbooks in labor and public economics and other fields (Moffitt, 2003). In addition, the 1970s IMEs led the way to widespread use of randomized trials to obtain credible evidence on the impacts of policies such as welfare-to-work, educational innovations (e.g. class size), job training and many others.

At the present time, there is widespread belief that – to paraphrase Solow – 'something needs to be done and could be done' to deal with high levels of inequality and the persistence of poverty and homelessness in otherwise wealthy societies. Many advocate a Basic Income (BI) or

Universal Basic Income (UBI), with an NIT design being a possible feature.[1] The probable effects of a BI or UBI on individual and family well-being, work activity, and marital stability – as well on as the incidence of poverty and extent of income inequality – continue to be assessed and debated. Although there are numerous BI pilot studies underway or recently completed from which much may be learned, the primary available evidence remains that from the 1970s IMEs. Reflecting this, many recent BI policy and survey papers rely heavily on this evidence (e.g. Widerquist, 2005; Marinescu, 2018; Hoynes and Rothstein, 2019; and Green et.al, 2023).

However, how reliable is this evidence? Economists have learned a great deal about the design and implementation of large-scale social experiments and the analysis of experimental data over the past 50 years. Recent papers by Price and Song (2018) and Riddell and Riddell (2024)[2] conclude that random assignment was not successfully carried out for specific target groups in two US experiments (Gary and the Seattle component of 'SIME-DIME'), raising doubts about whether this also holds for two-parent families—our analysis documents the same problems in the two-head experiments of Gary and SIME. The Manitoba IME was shut down after the operational phase without funds for research and analysis. No final report was issued, and only a limited amount of (non-randomly chosen) data was subsequently digitized and made publicly available to researchers. The sole published study of Mincome labor supply effects used data that cannot be replicated and does not line up with the publicly available data or other key features of the research design.[3] Despite these important limitations, the Mincome evidence has frequently been cited in BI policy discussions.[4]

These and other issues discussed subsequently— raise doubts about the credibility of the current consensus view regarding NIT impacts as reflected in the official Final Reports, many research papers in leading academic journals, and widely cited survey papers.

---

[1] Although the terms BI or UBI are sometimes interpreted as implying an unconditional grant, as discussed in the next section two of the most ambitious basic income pilots undertaken by governments (in Barcelona and Seoul) test a NIT design and several serious BI proposals by academics also employ that feature.

[2] We emphasize that our earlier paper in the Journal of Labor Economics on single mothers was based solely on the Mincome *public use file*. While this paper came out in print in 2024, it was accepted in 2022 and first submitted in 2021. It was through our attempts to replicate Hum and Simpon's estimates for single mothers that we learned a) of huge amounts of missing data, and b) that some fundamental mishap with the data used in the 1980s must have occurred. The grant from the Social Science and Humanities Council of Canada allowed for the audit of the Mincome hard copy records, and compilation of previously undigitized data (and commenced April 2023).

[3] We discuss previous Mincome research in detail subsequently.

[4] For example, one article states: "Mincome was a staggering success. The fear that people wouldn't work if they didn't have to proved unsubstantiated; working hours dropped only one percent for men, three percent for married women, and five percent for unmarried women" (https://www.vice.com/en/article/you-cant-talk-about-robots-without-talking-about-basic-income).

It is also noteworthy that the 1970s-era experiments focused rather narrowly on the potential adverse consequences of a NIT – largely in the form of labor supply reductions, but also on possible negative impacts on marital stability, an issue that arose in the analysis of SIME-DIME, the last and most ambitious of the US IMEs. Unlike recent and ongoing BI pilot studies, little attention was given to examining potential benefits such as household well-being, family functioning, and financial security.

In this paper we attempt to clarify the lessons learned from the North American IMEs about impacts on two-parent families – the primary target group in these experiments.
In addition to a re-assessment, a key contribution is to incorporate newly digitized data from Mincome – data that was previously available only in hard copy form. Doing so allows for an audit of the Mincome experiment, increases the Mincome sample size by about 65%, and also makes available hitherto untapped data on happiness, marital satisfaction and household production. In addition to re-assessing labor supply and marital dissolution, our analysis provides evidence based on this new Mincome data on the effects of the NIT treatment on broader household outcomes such as subjective well-being, time use and sharing of tasks within the household.

The current consensus view of labor supply impacts is succinctly summarized in Marinescu (2018): "The NIT experiments had a negative impact on employment among treated families, but this effect was generally not statistically significant." [5] Our re-assessment raises serious doubts about this view of the labor supply response in the North American IMEs. Specifically, our *precise* ITT estimates for women are: declines in labor supply of 21%-22% along both the intensive and extensive margins in DIME (almost exactly replicating the final report for DIME); but then contrary to the consensus for Manitoba and New Jersey, we estimate declines in labor supply of 17-26% (extensive) and 32-36% (intensive). Furthermore, ITT estimates for men in Manitoba, and men in the Pennsylvania site of the 'New Jersey' experiment are almost identical to DIME of (6-10%) and statistically significant. Overall, contrary to the current consensus of no effect of the NIT in New Jersey and Manitoba for both men and women, we find remarkable consistency in point estimates and statistical significance across experiments, genders and countries.

[5] The World Bank (2020) similarly summarizes the consensus: "Overall, the experiments find evidence of no effects or moderate reductions in work participation…the only consistently negative and statistically significant result arises from the Seattle-Denver experiment… Mincome in Canada recorded a modest reduction in hours worked which is statistically insignificant."

What explains such a dramatic change in the labour supply evidence? The reversal in estimated responses for Mincome and New Jersey are due to: a) prior Mincome estimates were invalid for multiple reasons relating to data errors and incomplete data (which we carefully document through our audit), and b) prior New Jersey estimates were based on roughly 50% of the complete sample of households due to (by today's econometric standards) the unnecessary exclusion of many observations. The panel data sample size was reduced by a further one-third by focusing on the middle two years of the experiment.

Overall, therefore, the costs of a NIT in terms of a reduction in labor supply are far greater than a reader would take away from the current basic income literature. The potential benefits, however, are also not well understood in the case of the North American IMEs. Based on previously unavailable data from the Manitoba experiment, we estimate that treatment group families report higher levels of happiness, marital satisfaction, and superior outcomes relating to the allocation of household tasks. The latter are also correlated with a reduced probability of marital separation, suggestive of long-run positive effects of the NIT offer on marital stability. We also find, for women, increased participation in social activities. The ITT estimates are particularly large for women with much more modest impacts for men. For instance, women randomly assigned to the treatment group reported a 27% increase in both satisfaction and agreement with the allocation of household tasks.

Finally, relating to household matters, we also show that the controversial result from SIME-DIME that the NIT offer caused higher divorce rates can be solely attributed to SIME (in which random assignment appears not to have been achieved). There is no evidence of an effect of the NIT offer on the probability of divorce in DIME or Manitoba.

2. Basic Income Literature

Interest and research on a basic income (BI) or universal basic income (UBI) continues to grow. Numerous books, articles and policy studies by both proponents and skeptics have been devoted to the topic, and governments, research institutes and wealthy individuals have responded by initiating pilot projects to assess the feasibility and impacts of a BI/UBI.[6] In the U.S. alone there are about 30 BI pilot studies with a randomized control trial (RCT) design

[6] Recent and widely cited books include Forget (2018), Green at. al (2023), Haagh (2019), Lowrey (2018), Murray (2016), Van Parijs and Vanderborght (2017) and Yang (2018).

underway or recently completed – many sponsored by local governments and carried out with the assistance of organizations with expertise in social experiments (see, e.g., https://guaranteedincome.us/). Government initiatives in other developed countries include those in Canada, Finland, Italy, South Korea and Spain.[7] Review papers on this rapidly growing policy area include Widerquist (2005), Marinescu (2018) and Hoynes and Rothstein (2019) for developed countries and Banerjee et al (2019), Hanna and Olken (2018) and McGuire et al (2022) for developing countries.

The term Basic Income is used to describe a diverse set of policies intended to reduce poverty and inequality. Typically, a BI refers to a cash benefit paid to recipients at regular intervals (e.g. monthly). These payments may replace parts of the existing social safety net, or supplement the income support system. Proposals differ, however, on several key dimensions. One is whether the BI is universal, i.e. received by all families, or income-tested, and limited to low-income families or paid to all but taxed-back according to family income. Another is whether the benefits are unconditional or conditional, e.g. requiring participation in the workforce or enrollment in education. While much discussion revolves around a basic income being 'universal', several authors question whether a UBI is fiscally realistic (Hoynes and Rothstein 2019; Green et al 2023; OECD 2017).[8] Many recent studies, especially those in developing countries, but also those in developed countries during the Covid pandemic, examine unconditional cash payments targeted on low-income families (e.g. Haushofer and Shapiro 2016, Londono-Velez and Querubin 2022, Jaroszewicz et al 2024, Pilkauskas et al 2023). Being targeted on low-income families, eligibility is income-tested but benefit payments typically do

[7] The Finish BI experiment operated from 2017-18 and focused on unemployment insurance recipients, many long- term unemployed (Verho et. al. 2022). In Canada, the province of Ontario was the first to introduce a BI pilot project (https://www.ontario.ca/page/ontario-basic-income-pilot) but it was cancelled after a change in government. British Columbia appointed an Expert Panel on Basic Income that carried out extensive consultations and research and recommended substantial changes in existing income support programs rather than a BI program (Green, Kesselman and Tedds, 2020). Prince Edward Island's pilot recommended a Basic Income program funded principally by the federal government (https://www.gbireport.ca/). In 2019 the Italian government introduced a BI referred to as "Citizen's Income" (https://www.oecdbetterlifeindex.org/countries/italy/) that was replaced in 2023 by a less generous and more restrictive policy by the current government. In Spain the city of Barcelona implemented a BI experiment B-MINCOME between 2017 and 2019 (Riutort, Lain and Julia, 2023). The Seoul Stepping Stone Income Project, a three-year RCT with an NIT research design, began in 2021. See https://seoulsafetyincome.welfare.seoul.kr.

[8] For example, based on their simulations for the Canadian province British Columbia, Green et al (2023 p. 163) conclude that "UBIs are so much more costly than IBIs (more than twice as costly to achieve the same level of poverty reduction) that it is hard to conceive of them as a reasonable policy choice". In the US context Hoynes and Rothstein (2019) conclude that "A pure UBI (providing a set benefit to all regardless of income, age, etc.) funded to meet basic needs for a household without earnings would be extremely expensive, about twice the cost of all existing transfers in the United States. Funding it would require substantial new revenue." Analysis by the OECD for numerous European countries reaches similar conclusions (OECD, 2017). Hanna and Olken (2018) also find that income-tested BIs strongly dominate UBIs in the developing country context.

not differ between those just below and those far below the eligibility cut-off. Implementing such a policy on a wide-spread basis would face major challenges. For these reasons, interest in income-tested policies including, in particular, a NIT remains strong. Indeed, it is noteworthy that current large-scale, government-funded experimental pilots designed to reduce poverty and inequality — such as Barcelona's "B-MINCOME" and Korea's "Seoul Stepping Stone Income Program"—employ a NIT design.[9] Clearly, a NIT-style basic income remains an important policy to understand.

While generally the basic income literature from developing countries is omitted from the wealthy country literature (e.g. Hoynes and Rothstein 2020, Marinescu 2018), we include some discussion of the former here given that the developing country literature has focused on a broader set of outcomes such as financial security, subjective well-being and health. We find it also informative to contrast the developing country results with recent US randomized control trials testing unconditional cash payments.

In the literature on unconditional cash transfers, the experimental evidence is mixed. Most experimental papers are in developing countries and find positive effects on a wide variety of outcomes. Over the last decade this literature has exploded —see Banerjee et al 2019 and McGuire et al 2022 for reviews. Rather than providing an exhaustive review we focus on (i) more recent studies contrasting selected experimental evidence in developing countries with recent US evidence, and (ii) those that examine outcomes similar to ours.

Haushoffer and Shapiro (2016) find large positive effects on subjective well- being and financial well-being/security (food security in particular) in Kenya. However, in another Kenya experiment that compares an unconditional cash treatment with free health care of the same value, Haushofer et al (2019) find no effects of the cash payment on health outcomes and various measures of subjective well-being. Handa et al (2018) find positive effects on financial well-being (in particular, measures of food security) in a government-backed experiment in Zambia. Banerjee et al (2020) examine an experiment of unconditional payments during the pandemic in Kenya, and find positive effects on measures of financial well-being (such as 'experiencing hunger') and health outcomes. An experimental study from Columbia examining cash payments also made during the pandemic found positive (although small) effects on various measures of

[9] Also, building on Boadway et. al. (2018), all concrete BI proposals in the Canadian academic literature involve an NIT design. Green et. al. (2023, chapter 6) review these proposals.

financial well-being/security such as making a loan payment (Londono-Velez and Querubin 2022). Overall, the experimental results from unconditional cash payments in developing countries is most consistent with positive effects on financial security with somewhat mixed but largely positive effects on subjective well-being and health outcomes. Although this literature has been more focused on outcomes such as health and subjective and financial well-being, Haushofer and Shapiro also examine labor supply impacts and find no evidence of negative impacts, a finding that is consistent with studies by Ardington et.al. (2009) and Banerjee et.al. 2017.

Conversely, three recent experimental studies carried out in the United States during the first year of the COVID pandemic find no effects on any of a large set of pre-registered outcomes. Specifically, Pilkauskas et. al. (2023) found no effect of a $1000 one-time payment on financial hardship, mental health or household outcomes comparable to ours including partner relationships (in one of the few studies to examine such an outcome). A follow-up study by these authors (Jacob et. al. 2022) also finds no impacts on pre-specified outcomes such as material hardship and mental health in the full study sample as well as among a very low-income sample. Similarly, Jaroszewicz et. al. (2024) found no effects of either a $500 nor $2000 one-time payment on financial well-being/security, health, or subjective well-being (all measured as indexes of multiple questions). We note that, generally (with some notable exceptions), this experimental literature on unconditional cash payments (the US studies in particular, and some of those from developing countries) examine outcomes over a much shorter time horizon than was possible with the NIT experiments. Also, some studies (in particular, the three US experiments) examine one-time payments [10] whereas other studies from developing countries along with both the original NIT experiments and current NIT pilots such as Barcelona and Seoul involve regular payments over multiple years (three to five years in the case of the North American IMEs, 3 years for Seoul, 2 years for Barcelona). The null results may also be due in part to the substantial US government assistance provided during the early pandemic period.

Perhaps most relevant to this paper – and an important exception to the short-duration studies discussed above -- is the study by Vivalt et.al. (2024) that provided an unconditional cash grant of $1,000 per month to low-income individuals (not families) in two US states for a period

[10] We note that this feature varies considerably across the developing country literature; indeed, Banerjee et al (2020) test three treatments, a one-time lump sum payment versus two long-term payment streams.

of three years, the same duration as the North American NIT experiments. Enrollment in this study began in October 2019 and continued until October 2020: thus, most of the experiment took place during the pandemic period, as was the case with the short-duration studies by Pilkauskas et. al. (2023), Jacob et. al. (2022) and Jaroszewicz et. al. (2024). An important feature of the Vivalt et. al. (2024) study is the attention paid to employment-related outcomes. They find evidence of labor supply reductions on both the extensive and intensive margins – a reduction in labor force participation of 2.0 percent for participants and in weekly hours worked of 1.3 to 1.4 hours per week. An interesting finding is that participants' partners reduced working hours by a comparable amount.

3. NIT Design and Implementation

A key feature of the North American NIT experiments was the Conlisk-Watts assignment model for allocating families to treatment plans.[11] Prior to random assignment, families were stratified by family type (two-parent families, single mothers with dependent children, and, in the Canadian case, single men and women); race (in Seattle and Denver), program length (SIME and DIME); location (in Gary and Mincome); and 'normal income' levels.[12] In New Jersey where there was only one family type, stratification was by normal income levels. Each stratified sample was offered treatment plans that combined different guarantee levels *G* and implicit tax rates *t* in an attempt to facilitate estimates of the responsiveness of families to NIT plans with different incentives.

An important consequence of the Conlisk-Watts assignment model is that for the sample as a whole there is non-random assignment to treatment and control groups. Rather, random assignment took place within combinations of the experimental stratifications noted above that were adopted for a particular experiment. For two-parent families, this includes normal income in all experiments. In order to obtain unbiased estimates of treatment effects it is therefore necessary to control for the appropriate stratification categories as well as interactions among

[11] This model, first described formally in Conlisk and Watts (1969), is designed to optimize the allocation of families with different pre-treatment income levels to the various treatment plans, taking account of the overall budget for the experiment. Pure random assignment of families to alternative treatment plans would result in some low-income families being offered very generous (high guarantee G, low tax-back rate t) treatment plans – resulting in very expensive observations. Essentially this assignment model reduces the likelihood that families with very low pre-treatment income (and raises the likelihood that families with relatively high pre-treatment income) are enrolled in generous treatment plans relative to what would occur under pure random assignment.

[12] Normal or permanent income was computed from pre-treatment surveys discussed subsequently.

these categories (see, e.g. Athey and Imbens, 2017). We use the term 'experimental cell' to refer to the level at which random assignment takes place. The number of experimental cells varies substantially across the NITs. As the only data digitized for Mincome is the Winnipeg site, the experimental cells for two-parent families in Winnipeg consist of only the 4 normal income categories. DIME has the largest number of stratification groups; even within the two-parent family category there are 5 income categories, 3 races (Black, White and Hispanic), and 3 durations (3-years, 5-years and 20-years). One consequence of this model is that sample sizes are small for individual experimental cells. Anther consequence is that there is unbalanced allocation to treatment and control groups – the sample size of the control group is typically much smaller than the treatment group (approximately 60-40 in most cases). Perhaps the most important issue to note is that the early literature did not always control properly for these stratifications.[13]

4. Data, Balance and Attrition

(a) The New Jersey Experiment

New Jersey was the first of the North American IMEs. There were three sites in New Jersey and a fourth in Pennsylvania. Enrollment was sequential, beginning with a small 'pilot site' enrollment in Trenton in August 1968 (155 families) followed by much larger enrollments in Patterson-Passaic (January 1969), Jersey City (May 1969) and Scranton, Pennsylvania (October 1969). 1357 families were enrolled. The experiment lasted three years at each site.

Unlike subsequent IMEs, New Jersey focused solely on two-parent families – more precisely, families with an able-bodied male between 18 and 58 years of age and normal income not more than 150% of the poverty line. In order for differences between treatment and control families to reflect the effect of the negative income tax alone, it was considered desirable to conduct the experiment in a state in which male-headed families were not eligible for any welfare assistance (Pechman and Timpane, 1975). New Jersey was chosen in part because it did not have a plan under the federal AFDC-UP program that extended financial assistance to unemployed fathers. However, on January 1, 1969 New Jersey introduced an AFDC-UP plan for which most families in the experimental sample were eligible. The New Jersey AFDC program,

[13] Specifically, the early literature often simply included fixed effects for each separate stratification category. However, the researcher needs to include a dummy variable for each experimental cell (or a full set of interactions between all stratification variables).

including the AFDC-UP component, was one of the most generous in the US (Garfinkel, 1977). Thus, for the Trenton 'pilot site' there was a major change in the policy environment after the first quarter, and for the other three sites the new policy regime required some changes in planned operating procedures.[14] One key change was to require families to choose between welfare and the NIT each month, a policy adopted subsequently in Gary, SIME-DIME, and Mincome. Thus, aside from Trenton, the interpretation of the experimental estimates in New Jersey is comparable to the other experiments we consider in which the NIT and welfare co-exist (additional details are presented in Online Appendix 5).

The labor supply analysis in the New Jersey Final Report (Watts and Rees, 1976) restricted attention to a 'continuous husband-wife subsample' ("in-tact") – 693 families that remained as married couples throughout the experiment and who responded to the pre-enrollment survey, 12th quarterly interview, and missed not more than 5 quarterly interviews of which not more than two lapses could be consecutive. Focusing on this non-representative sub-sample limited the analysis to about 50% of families enrolled. In addition, labor supply analysis was carried out for only the middle two years i.e. quarters 3-10, the reasoning being that in the first 6 months the families were getting adjusted to the NIT and in the final two quarters they might be adjusting to the end of the NIT. The panel data sample size was thus reduced by an additional one-third by limiting the analysis to the middle two years of the experiment. We note that dropping attritors was common at that time—the SIME-DIME Final Report made a similar sample restriction.[15]

Our analysis uses the New Jersey 'Cross-Site' file.[16] We use all 12 quarters of data, and (as with all of our analysis) include year dummies to deal with the potential effects of adjusting to the beginning and end of the experiment, among other factors correlated with time. We report results for all four sites pooled in the main paper.[17] It may be appropriate to exclude the Trenton

[14] Although the introduction of a generous AFDC-UP policy was unexpected by the organizers, it is worth noting that a less generous state welfare program for which two-parent families were eligible existed at the time the N.J. experiment began. Also, the generosity of the AFDC-UP program was short-lived – benefits were substantially reduced in July 1971.
[15] The SIME-DIME Final Report required households to have remained in the data for the first three years of the experiment, and thus attritors were dropped for the three-year stratification, but not all attritors were dropped for the five-year program.
[16] The only data currently available for New Jersey is the cross-site file; note (as stated subsequently) that for DIME, Gary and SIME we use the full monthly files (which in addition to monthly variables on labor market activities, take-up, etc. have more information available relative to the cross-site file).
[17] While details are thin in the New Jersey Final Report (and related studies cited here), for a variety of reasons we suspect that the 'site' should be considered an experimental stratification in the 'New Jersey' experiment. We therefore include full interactions between income cell and site for the pooled sample results included in the main paper (along with the robustness checks in the Online Appendix).

'pilot site' because of the January 1 1969 policy change.[18] There may also be reason to report separate findings for Pennsylvania given different state laws. In the Online Appendix—and discussed in the Results section below—we therefore also present results for three additional samples (full sample excluding Trenton; full sample excluding Trenton and Pennsylvania; and Pennsylvania separately). However, we do not exclude the many observations dropped in the early research including (i) attritors, (ii) non 'in-tact' households (e.g., households with a separation or divorce among other cases); or (iii) participants who failed to respond to an arbitrary number of quarterly interviews. Finally, we note that there is no data on take-up for the New Jersey experiment.

Baseline summary statistics for New Jersey are discussed together with those for Mincome and DIME subsequently.

(b) The Mincome Experiment[19]

Mincome was a joint federal-provincial initiative carried out in Manitoba in 1974-78. There were three sites: Winnipeg, the rural dispersed sites and the non-experimental 'saturation site'—the town of Dauphin in which all low-income families were eligible. We ignore the non-experimental Dauphin site as well as the rural sites (due to very small sample sizes).[20]

Mincome had a budget fixed in nominal dollars at a time of high inflation. Although there was substantial interest in poverty reduction at both the federal and provincial levels when Mincome was introduced, the recessions and stagflation of the early 1970s together with changes in governments at both the federal and provincial levels resulted in different priorities. As a consequence, when the 1974-78 operating phase ended Mincome was shut down without any funds for research and analysis. No final report was produced, and the survey and payment records remained mainly in hard copy form, subsequently transferred to the National Archives.

In 1981 the federal government provided some funding to restore the Mincome data and promote its use. By 1983 some of the data from the Winnipeg site had been digitized, together with detailed codebooks, and was made available to researchers. However, by this time interest

[18] As well, there is reason to believe that certain decisions were made or changed during the first year of operation in Trenton that result in the Trenton data being "less reliable than the data from the other sites" (Skidmore, 1975). We find that Trenton alone passes balancing tests although the sample size is extremely small.
[19] This section provides a brief overview of Mincome. More detail is available in the various technical reports and studies referred to in Simpson, Mason and Godwin (2017).
[20] We note as well that the survey data has never been digitized for Dauphin or the rural sites; only the administrative data (see subsequent discussion) has been used in the case of Dauphin for labor supply, and no research at all exists for the rural site.

in the guaranteed income policy had waned so only limited research was carried out. As a consequence, Mincome remains dramatically under-researched relative to the US NIT experiments. Indeed, until our recent paper on single parents (Riddell and Riddell, 2024) only one published study of labor supply effects using the Mincome survey data (i.e., from the randomized Winnipeg experiment) – that of Hum and Simpson (1991) – had been carried out.[21] We have been unable to replicate the results of this study and the published study does not provide sufficient detail about how the data were processed to be helpful.[22] Neither the data used by the authors or their code are available. Of particular note, the sample sizes in Hum and Simpson do not match the sample sizes in the Mincome official documentation (nor the current public use file[23]), and are inconsistent with the assignment model. We outline these issues in more detail, and provide an accounting of the public use microdata sample sizes relative to those reported in Hum and Simpson in the Online Appendix (see Section 3). We also document a crucial data error in the public use file that could explain the discrepancies. Based on our doubts about the validity of their sample and the inability to replicate their findings, our view is that no credible evidence on labor supply impacts for two-parent families in Manitoba currently exists.

Of course, another possibility is that the current public use file (and official Mincome documentation) is inconsistent with Hum and Simpson because records were missed when the data was digitized. Moreover, a variety of additional data was never digitized due to the budget issues noted above. These data are discussed in more detail below, but briefly, include: a) non 'in-tact' households such as households with a marital separation (even if not formal divorce) as well as other 'discontinuities' in a household head (such as death, imprisonment, temporary separation—for instance due to employment—from the household, etc.); b) a variety of modules or specific variables for the digitized households were omitted or never defined in the available documentation; and c) the Winnipeg Supplemental Sample. As noted subsequently, the addition of a) and c) would result in an increase in the sample size of 65% in addition to addressing the

[21] Results from this study are also reported in Hum and Simpson (1993) that surveys evidence from the US and Canadian income maintenance experiments. Calnitsky and Latner (2017) carry out a non-experimental analysis of the extensive labor supply margin using administrative data (see our discussion below) from the Dauphin site, which was not randomly assigned.
[22] Notwithstanding the more fundamental issue of the underlying data, it is also important to clarify that Hum and Simpson (1991; 1993) pooled husbands from two-head households with single men with no dependent children for estimation of labor supply effects. Mincome was the only IME to conduct an experiment with single adults with no dependent children. The latter may not respond to a NIT offer in a similar way to husbands with dependent children. Moreover, randomization was conducted differently for single adults with no dependent children; the income cells were distributed very differently, and thus it is inappropriate to pool the two experiments. In any event, the results for men from Hum and Simpson are not comparable to the US results, a point missed by the previous literature.
[23] Available from the University of Manitoba Libraries.

potential sample selection bias of excluding non- 'in-tact' households. Given these various issues, we hand-collected all missing data, as well as the family identification records of all households from the original hard copy records of the Winnipeg experiment at the Archives of Canada. Importantly, doing so also allows for an audit of the original Mincome data. By re-constructing the household identification records from the original files, we can establish what the true sample size is, as well as add missing data. With respect to the audit component of this study, we ultimately find slippage of less than 20 households between the records at the Archives of Canada and the current public use file.[24] A summary of our audit of the Mincome data with respect to past research including our final sample size is contained in the Online Appendix (see section 4).

The Mincome data collection process was built on the US experiments, and thus has a similar basic structure.[25] However, some differences exist largely due to the budgetary issues noted above. Overall, the Mincome data consists of two pre-random assignment surveys—the baseline survey (referred to as "Minc1") and the enrollment survey—and nine post-random assignment surveys (referred to as "Minc4").[26] However, monthly data (based on weekly hours reports from the jobs records) was never created in the Canadian experiment, and thus the panel structure is 9 post random assignment data points (see further below).

Two new, never digitized, data sources are available for the Winnipeg experiment in hard copy form at the Archives of Canada (Winnipeg office). The first is the "husband-wife" module—a unique feature of Mincome that included questions relating to happiness, marital satisfaction, satisfaction and agreement with household duties, and involvement in social activities such as church-going. This module was first incorporated in the pre-random assignment enrollment survey—which we use in our balancing tests—and then administered again at the 5th

[24] There are nine cases of households that appear in the public use data but for which no record can be found at the Archives (four of these are consecutive family IDs and thus likely are in a single box that has gone missing), and 7 cases of records that are at the Archives but do not appear in the public use microdata file. One final possibility that cannot be ruled out is that some records were lost between the original data collection and storage at the Archives of Canada; however, files were stored chronologically by family ID number not treatment status, and thus the likelihood of hundreds of control group records being lost seems remote (see Online Appendix, Section 4).

[25] In particular, with respect to labor supply, households completed weekly hours worked reports for each reported job.

[26] The public-use Minc4 file contains all labor market data for the post random assignment period. Non-economic data is available in a separate file referred to as Minc6; however, large parts of this data are useless because no codebook was ever created.

periodic. (approximately 2 years later). The Online Appendix (section 2) contains details on the questions asked and data availability.[27]

The third source of additional data is the Winnipeg Supplemental Sample. Mincome's original sample size was notably smaller than even New Jersey. In the early stages of the Mincome experiment there were concerns about low take-up, early drop out (between the baseline and enrollment surveys), and that generally non-participation could result in insufficient observations in some treatment plans (especially in the less generous NIT plans). An additional sample—labelled the "Supplemental Sample" of approximately 260 Winnipeg families was therefore added shortly after the 'original' Winnipeg experiment. However, these data were never digitized. Interview dates for the first (post random assignment) periodic survey in the original Winnipeg experiment were staggered from February to September 1975 while the Supplemental Sample was staggered from January to April 1976. [28]

Digitizing all missing data for the Winnipeg site (non 'in-tact' original Winnipeg households plus the Supplemental) results in a sample size of 821 couples – an increase of 65% relative to the previously digitized (i.e., current public use) Winnipeg data.

(c) Seattle-Denver

We group Seattle-Denver (SIME-DIME) together for purposes of this section as the data structure is identical. Indeed, it has been common in the literature to refer to Seattle and Denver as essentially one experiment; previous analysis has pooled the two together, and much of the literature refers to the US NITs as consisting of four experiments (New Jersey, Rural, Gary and SIME-DIME). However, there are important differences between the Seattle and Denver experiments so we analyse them separately.[29] Because SIME-DIME has received considerable attention in the academic and policy literature we restrict our discussion to key points about SIME-DIME's data structure that have received less previous attention.

---

[27] Other than labor market information, virtually all questions in Mincome were not asked in every periodic. For example, the happiness question was asked three times. The public use file does have some data (in the Minc6 file) from the husband-wife module, but the specific questions asked, and the codebook for this information were never digitized.

[28] These households were principally drawn from the original Winnipeg screening sample. Given available budget only a fraction of households completing and passing the screening survey were approached for the original sample. When the decision was made to expand the sample more households were added using the same sampling criteria and same experimental stratifications (income cells). Thus, the Supplemental Experiment can be considered akin to a later entry cohort in a staggered experiment. The overall staggering of the Winnipeg experiment (including the Supplemental) was roughly the same as Seattle and Denver, and shorter than New Jersey (if we include Trenton).

[29] In particular, along with occurring in different states, SIME and DIME used different experimental stratifications; had different staggering of entry; and occurred at different points in calendar time.

We use the SIME and DIME 16th Monthly Composite Principal Person Files. The original SIME-DIME data collection was similar to Mincome in that pre-random assignment surveys were collected followed by post-random assignment periodic surveys (also roughly 4 months apart). However, the data that was digitized for public use, the SIME and DIME 16th monthly composite files, have important differences from New Jersey and Mincome. First, the labor market information was primarily collected from job start and end dates; from those, a 72-month panel was constructed. Second, while New Jersey, Mincome and SIME-DIME had staggered entry, SIME-DIME collected the 72 months of data over the same calendar time period. Thus, cohorts differed in the number of months that constitute pre-random assignment data. Moreover, SIME-DIME had both 3- and 5-year programs, and thus for the 3-year program, there is also post-experiment data.[30] The number of months of post-experiment data also varies by entry cohort. This contrasts with Mincome and New Jersey where no information on participants is available after the experiment ended.

Note the pre-random assignment period varies from 10 to 22 months. In our balancing tests below, we focus on months 1-9 as these months are pre-random assignment observations for all individuals.

A non-trivial issue that arises in the SIME-DIME data is the extent to which a variety of family types are, in fact, 'two-head households with dependent children'.[31] A close look at the data reveals that there are a non-trivial number of observations that, at some point over the *pre-random assignment period*, do not satisfy this definition, and thus we exclude them from the sample.[32] In particular, we drop the following cases: households where the head is an adult child

[30] Note that DIME also had a small 20-year program (later abandoned) that we exclude from our analysis (see below for sample construction discussion).
[31] This issue is less severe in New Jersey, perhaps because their eligibility requirements were different than SIME-DIME, Mincome and Gary. Finally, we refer in the text to DIME, but all of the sample selection issues raised in the paper apply to SIME, and relative counts of observations excluded are very similar.
[32] There are also a variety of households in the original 16th Monthly Composite data file that never make it to the enrollment survey (similar to Mincome as noted in the text). For general interest purposes—and also in reference to our New Jersey analysis—it is possible in the case of SIME-DIME that some of these records were deleted when the "cross-site" file was created. Specifically, there are 606 observations (303 households) excluded that never responded to the stratification requirements (i.e., income levels) or 'dropped out' prior to the enrollment survey.

or other relation, and households with no dependent child (i.e., the 'children' are adults).[33, 34] Ultimately, our sample at the time of random assignment consists of 1503 households.[35]

Unfortunately, it is exceedingly difficult to reconcile the sample sizes of the early SIME-DIME studies with the public data. The special issue on SIME-DIME in the *Journal of Human Resources* (Spiegelman and Yaeger, 1980) reports a sample size of 1597 households for DIME, and Robins and West (1980) report a SIME-DIME sample size of 2928 households, which roughly amounts to 1654 for DIME alone.[36, 37] We can find no documentation that explicitly declares a starting sample nor what deletions were made.

(d) Baseline Summary Statistics

Baseline assignment summary statistics are shown for Mincome, DIME and New Jersey respectively in Tables 1-3 (see row labeled 'Mean Dependent Variable').[38] In interpreting these keep in mind that the pre-random assignment surveys differed in the time period covered (see Notes to each table) and, in the case of Denver, even differed across individual families as some had longer pre-random assignment periods than others. Also, the Mincome baseline survey covered the pre-random assignment year 1974, so isn't affected by seasonal employment whereas the Denver survey is monthly and New Jersey quarterly.

---

[33] As examples of the former, this tends to consist of cases where the (likely at the screening stage) original heads are older and the adult child becomes head for purposes of NIT enrollment/eligibility over the pre-random assignment period, as well as cases where a spouse or partner had applied for eligibility but was still waiting for the decision. These cases amount to 177 observations at the household level excluded.

[34] We also treat as attritors (i.e., drop the observations after the change) those observations where the household's treatment status was changed. Unusually, there are cases in DIME where a control group household was switched to a treatment household during the experimental (i.e., post random assignment) period (incidentally almost all of the 20-year DIME program were former members of the control group switched after the time of random assignment). This does not affect the sample construction at the time of random assignment.

[35] This is our sample size at the time of random assignment (e.g., used for balancing tests). The sample size is smaller for labor supply estimation due to drop out between random assignment and the first month of the experimental period as well as non-response on outcome variables.

[36] Across various starting points, DIME accounts for about 56-57% of the SIME-DIME total.

[37] The SIME-DIME Final Report (SRI International, 1983) states a sample size for purposes of analysis (i.e., not at the time of random assignment/enrollment) of 1911 for 'husbands' for SIME-DIME pooled—amounting to approximately 1070 for DIME alone—and 2043 for 'wives'—amounting to about 1140 for DIME. However, the Final Report dropped attritors for the labor supply estimation and does not report the number excluded. Our 'husband' and 'wife' sample sizes are 1503 for women and 1459 for men. Differences between men and women are largely due to non-response on outcome variables.

[38] The hours worked variables are not comparable across surveys, in particular the baseline survey (used in our balancing tests, see Table 1) definition differed from the both the enrollment survey (which we do not use for labor supply for this reason) and the post random assignment periodic surveys. The main differences were in the way casual work (often referred to as 'odd jobs') and paid hours not worked were incorporated into the hours worked variable. Thus, we use hours worked from the baseline in our balancing tests but then we use the control group mean for the experimental period for estimating magnitudes of the treatment effects (and present both summary statistics in the tables).

Despite these differences, some common features are evident. Male employment rates are similar in each IME, ranging from 84% in DIME to 88% in Manitoba. Male hours worked are also broadly similar once the time period covered by the surveys is taken into account. Wives' employment rates and hours worked are much lower in each IME, with the gender gap being largest in New Jersey and similar in Denver and Manitoba. Hours worked by men in New Jersey is more than 7 times greater than that of women, versus about 4 times greater in DIME and Mincome. Similar gaps exist in employment rates – between 2 and 3 times greater for husbands in DIME and Mincome but almost 6 times greater in New Jersey.

Men tend to be 2 ½ to three years older than women in each IME, with New Jersey having the oldest and Denver the youngest participants. Welfare receipt is almost identical between men and women as expected for a program that is family based, a bit greater in New Jersey at 17-18% and lowest in Mincome at 11-12%. Almost two-third of participant families in Mincome and Denver have children less than school age, versus 47% in New Jersey.

(e) Balancing Tests and Attrition

Although it is now common to verify that the experimental sample is appropriately balanced, tests for balanced samples do not appear to have been reported in the original NIT literature.[39] Recent studies by Price and Song (2018) and Riddell and Riddell (2024) find that balance was not achieved for specific target populations in the Gary and Seattle experiments (single parents in Seattle and Gary in Riddell and Riddell, and two-parent families with at least *two* children in Seattle in Price and Song). This raises doubts about the likelihood that randomization was successful for two-parent families, and thus that these experiments will yield unbiased estimates. Both cities experienced quick and dramatic downturns in the dominant industry (steel in Gary and aerospace in Seattle) around the time the experiments began, events that appear to have affected the treatment and control groups differently due to different timing of experimental in-take dates (see SRI International (1983, Vol II, Chap. 3) and Riddell and

[39] Several authors (e.g. Keeley and Robins 1980; Robins and West 1980) point out that there were pre-experimental differences in labor supply of treatment and control families, as well as different trends in work activity during the experiment. However, as they also note, such differences are expected given the nature of the assignment model. By "balancing tests" we refer to tests for treatment-control differences controlling for stratification categories (including normal income). Under such tests failure of random assignment could alter our interpretation of the treatment effects as causal impacts.

Riddell (2024) for details on problems encountered in enrollment and assignment to treatment in Seattle.)[40]

Appendix Tables A1 and A2 present tests of balance in pre-treatment labor market outcomes for the two-parent family samples. These results confirm our doubts about the two-parent family samples in these cities. In Seattle, treatment group males had lower earnings and higher welfare receipt while their female counterparts had lower earnings and hours worked. These treatment – control differences are highly statistically significant. In Gary, men in the treatment group reported higher employment levels and hours worked, and females had lower levels of welfare receipt. The magnitude of the earnings differences in SIME (roughly 50% gap), and welfare receipt in Gary are also worrying.

In contrast, our two parent family samples appear balanced in Mincome (Table 1), Denver (Table 2) and New Jersey (Table 3) based on demographic characteristics and labor market outcomes. In Mincome there are no statistically significant treatment – control differences prior to beginning of treatment, and in Denver and New Jersey there is only one difference that is statistically significant (at the 10% level, and small in magnitude).

As noted above, the Mincome husband-wife module—which contains most of the key questions for non-economic outcomes (except 'happiness')—was administered pre-random assignment at the enrollment survey. This allows us to conduct balancing tests on the various marital and household time allocation measures. These results are presented in Table 4, and we see that in the enrollment survey there are no statistically significant differences between treatments and controls for any outcome.

We also check for evidence of non-random attrition, another potential source of bias. Attrition in the U.S. NITs has been discussed previously, and there is evidence of non-random

[40] Not surprisingly, the intake interviews for the two-head sample were conducted in the identical manner as for single parents. In SIME, the entire control group entered the experiment on December 1970 while the treatment group was unevenly staggered from October 1970 to October 1971 with 50% of the treatment group entering between August 1971 and October 1971. The recession in Seattle—due principally to the near collapse of Boeing—was described as "probably the sharpest and most rapid rise in unemployment of any major city since the depression of the 1930s." (Rainey et al 1973). The aerospace recession began in 1969, but the peak for unemployment rates occurred in the spring-summer of 1971 at 14% (compared to the national average of 4.5%)—which could explain the lower earnings levels and higher welfare receipt in the treatment group. As well, within the treatment group, there was uneven staggering of the experimental stratifications (in particular, the income cells). In Gary, enrollment was unevenly staggered from October 1970 to October 1971, and the collapse of the steel industry began in June of 1971. Reporting at the time indicating that the 1971 layoffs would amount to 25 000 individuals out of a town of population 175 000 (New York Times 1971). Unlike SIME, in-take was at least staggered for the control group, but there was still a 14-percentage point differential in entry rates between treatments and controls pre/post June 1971. Moreover, in-take was not distributed evenly over the stratifications; in particular, the urban/rural stratification was unevenly distributed over calendar time with rural households disproportionately entering the experiment later. Of course, other issues could have occurred with random assignment, but these calendar time shocks may explain the pre random assignment differences in labor market characteristics.

attrition in SIME-DIME pooled data (Ashenfelter and Plant, 1990). We perform the test outlined in Fitzgerald, Gottschalk and Moffitt (1998) by regressing pre-random assignment outcomes (labor supply and non-labor supply) on an attritor dummy (and other individual characteristics). If attrition is independent of potential outcomes, it should not be correlated with pre-random assignment outcomes.

Results are shown in Tables 5 (for Mincome) and 6 (for DIME and New Jersey). In the case of Mincome, for both males and females there are no statistically significant correlations between the attritor dummy and baseline measures of our key outcome variables including labor supply and the available household and marital satisfaction measures discussed above. Similarly, in Denver there are no significant correlations between the attritor dummy and hours worked and employment prior to random assignment. In the case of New Jersey for men, we do find some evidence—for hours worked although not the probability of working—of attrition bias. For SIME-DIME (which has received the most attention with respect to attrition), our findings suggest previous evidence of non-random attrition in SIME-DIME arises from the Seattle experiment – possibly for similar reasons underlying the unbalanced sample in SIME.

Due to the results of the balancing and attrition bias tests, the remainder of the paper will focus on Denver, New Jersey and Mincome where random assignment appears to hold and where we find little evidence of non-random attrition.

5. Results

We begin this section with our re-assessment of labor supply in Denver, New Jersey and Mincome. We then present our analysis of the hitherto untapped Mincome husband-wife module along with a re-assessment of marital dissolution.

(a) Labor Supply

The Online Appendix (section 1) summarizes the previous literature on labor supply responses of males and females in two-parent families based on survey articles by Robins (1985), Burtless (1987) and Hum and Simpson (1993) and the SIME-DIME Final Report (SRI

International, 1983).[41] Overall, as the representative quotes noted in the Introduction illustrate, with the exception of wives in Gary all point estimates are negative on both the intensive and extensive margins for men and women in both countries, but only those for DIME, SIME and SIME-DIME pooled are statistically significant. The lack of consistency in the estimated magnitudes of responses across experiments, ranging from small and insignificant estimates for men and women in Mincome to larger and precisely estimated results in SIME-DIME, makes generalizing from the experiments difficult.

Our estimated treatment effects for the three experiments are reported in Table 7.[42] For females in two-head households in Mincome our ITT estimates imply a 26% decline in hours worked – an impact that is statistically significant at 1% – and an estimated 17% decline in the employment rate. This estimated decline in hours of work is far greater than the negligible and statistically insignificant estimate reported by Hum and Simpson (1991), but similar in percentage terms to those in DIME. Likewise, our estimated annual hours reduction (144) is similar to the 103 hours for DIME reported in the SIME-DIME Final Report (SRI International, 1983) and our estimate of 101 hours for DIME. The SIME-DIME Final Report does not provide separate employment rate estimates for DIME and SIME. For men, our Mincome ITT estimate of a 123 annual hours reduction translates to about an 8% decline (and is statistically significant at the 5% level), and nearly a 5-percentage point decline in the probability of employment (a 5% decline). Hum and Simpson (1991) report annual hours point estimates of 17 hours, nearly zero in percentage terms (and statistically insignificant). Our Mincome estimates for men are almost identical to DIME.

As noted above, for DIME, we essentially replicate the DIME-only point estimate of the SIME-DIME final report despite uncertainty around sample construction, and not dropping attritors (which is consistent with our tests for attrition bias). The ITT estimates imply declines in hours worked and the probability of working of 21% and 22% respectively for women, and a 10% decline in hours and 6% decline in the probability of working for men. Our hours worked ITT estimates *in percentage terms* are modestly greater than the most frequently cited percentage

[41] The sole previous study of labor supply in Mincome (Hum and Simpson, 1991) pooled together men in two -parent families with single men, no kids (21% of their sample) so aside from data replication issues, there actually are no previous estimates for Mincome husbands from two-head households with kids as a separate group.

[42] As noted previously, there are three surveys annually in Mincome, quarterly data in New Jersey (due to using the 'cross-site' file), and monthly surveys in DIME. To ease comparison, we include a row that reports the size of the hours of work coefficient in annual hours equivalents.

effects for SIME-DIME pooled of 14% for women and 7% for men.[43] We also emphasize that our estimates are almost identical to Mincome for men, and broadly comparable for women.

The estimates for New Jersey also imply very different conclusions—at least for women—than the consensus view. Using the full sample, the ITT estimate of about 22 hours for women (equivalent to 86 annual hours) implies a labor supply reduction of 36% along the intensive margin. The 5-percentage point reduction in the probability of employment –almost identical given the standard errors to both Mincome and DIME—is a 32% decline along the extensive margin. As noted previously, the consensus results for New Jersey are that labor supply effects are not statistically significant for either gender. We note, however, that the female point estimates in the New Jersey Final Report (based on all experimental sites)—and implied percentage reduction for the intensive margin (25% reduction in hours worked according to estimates reported by Burtless (1987) and Robins (1985)) is broadly similar to our estimate of 36% for all sites.

For New Jersey men our ITT estimates imply no significant reduction in work activity on either margin, the sole finding in this paper that is consistent with previous estimates. For example, our estimated decline in annual hours (35) is statistically insignificant and similar in size to estimates in the survey papers by Robins (1985) and Burtless (1986).

In the Online Appendix (section 5), we present robustness checks for New Jersey based on the experimental sites included in the sample. Estimated impacts are somewhat larger for the main New Jersey sites (Jersey City and Patterson-Passaic) with estimates of 39%/36% (intensive/extensive) when the small Trenton pilot sample (N=146) is excluded, and 44%/41% when Pennsylvania is also excluded. The larger effect in percentage terms for New Jersey women is due to a much lower baseline employment rate relative to DIME and Mincome.[44] Finally, we note that our estimates for men in Pennsylvania are almost identical to the results for

[43] Several differences between our sample and those used in the previous literature make comparisons difficult. As noted previously, to the best of our knowledge, all of the published academic literature pooled the Seattle and Denver experiments into a single dataset. The only estimates for DIME alone are from the SIME-DIME Final Report which did not provide estimated means of hours worked for either the control group over the experimental period nor over the pre-random assignment period (thus making percentage ITTs relative to a benchmark difficult). For the intensive margin, the Final Report presents an estimated ITT of 103 hours for women—compared to our estimate of 101 hours—and 150 hours for men—compared to our estimate of 173 hours. Final Report estimates are for the second year of the experiment whereas ours are for all years. Well known review papers for the SIME-DIME pooled evidence are Robins (1985) and Burtless (1987). Burtless presents point estimates relative to pre random assignment summary statistics for SIME-DIME pooled—estimates of 7% for men (our DIME estimate is 10%) and 14% for women (our DIME estimate is 21%), which have been widely reported in the basic income literature. The notably lower effect reported by Burtless is due to a higher (pre-random assignment) hours worked benchmark (his calculated treatment effect is also 101 hours, see Online Appendix, Section 1, footnote 1).

[44] The estimated decline in annual hours is 86 in N.J. versus 101 in DIME and 144 in Mincome.

men in Manitoba and Denver at 9% and 6% for intensive and extensive margins respectively, and statistically significant.

Overall, our ITT estimates for New Jersey, Mincome and Denver suggest very different labor supply conclusions than the consensus results. The recent World Bank (2020) review on the North American NITs (echoed in many other reviews) was: "Overall, the experiments find evidence of no effects or moderate reductions in work participation… the only consistent negative and statistically significant result arises from the Seattle-Denver experiment." It is a fair assessment to state that, across many reviews of the NIT experiments, SIME-DIME was viewed as the outlier in terms of the magnitudes and precision of the estimated impacts. Our re-assessment reveals that DIME, if anything, is on the lower end of labor supply impacts for women, and almost identical to the Canadian experiment for men. To summarize, this re-assessment of ITTs stems principally from two changes: a full sample for New Jersey (as opposed to a 50% sample that dropped many observations for reasons that would not be dropped today and is less representative of the low income population), and a re-constructed, complete Mincome sample that –in addition to adding observations dropped for reasons similar to New Jersey—incorporates data not previously digitized that can be reconciled against the original hard-copy records.[45] Our estimates for DIME also do not restrict the sample to families that remained in the sample for the first three years of the experiment.

(b) Subjective Well-Being and Time Use

We turn now to an analysis of some possible benefits of a NIT-based basic income. Table 8 presents intent-to-treat estimates of the impact of the NIT offer on happiness, agreement and satisfaction with contributions to household production and satisfaction with the marriage, and

[45] In the Online Appendix (see section 6), we present all Mincome results for the current public use file. All of the estimates for women are almost identical (even with loss of precision, all are statistically significant at conventional levels). Labor supply estimates for men are similar qualitatively but just outside of conventional significance levels based on the public use file. *Thus, even labor supply estimates from the current public-use file for two-head households do not line up (for women, and point estimates do not line up for men) with Hum and Simpson (1991)*. We remind readers that Hum and Simpson included single men without children (24% of the sample) in their "male" sample along with husbands (from the two-head sample). Mincome was the only North American NIT that included the 'singles-no kids' population. This feature alone of Hum and Simpson makes it likely not comparable with the existing literature. However, we note that even if we try to replicate their results with the public use file and include single men in the sample, we still cannot replicate either their point estimates or sample sizes.

our proxy for social activities (church attendance). For both men and women, we find an increase in happiness of about 5 percentage points, or around a 6% increase in happiness.[46]

On household production, both women and men in the treatment group report higher levels of *satisfaction* with their spouse's contribution to these activities. For women in particular, this is a sizeable estimate given that baseline satisfaction with their husband's help around the house is not as high as most of these indexes. The control group mean indicates that 57.4% of women are 'satisfied' with their husband's help; the ITT of 15.5 percentage points implies about an 27% increase. On *agreement* with household duties, the results are similarly suggestive of a more beneficial intra-household allocation of housework for women (ITT estimates also indicate a 27% increase for women). The evidence is far more modest on the husband's side.

The estimated impacts on social activities—as proxied by church attendance[47]—suggest that households at least partly re-allocated labor market time towards additional personally rewarding activities beyond household production and pure 'leisure'. The results differ by gender with notable increases in church attendance only for women (roughly a 20% increase). This finding on church attendance seems consistent with previous evidence on labor supply, which indicates larger reductions in hours of work and employment for women than for men.

(c) Satisfaction with Marriage and Divorce

The potential impact of a NIT on marriage dissolution has long been a controversial aspect of the US NIT literature. The concern arises because the NIT may increase the combined income of the spouses when single relative to being married. In a series of papers, Hannan, Tuma and Groeneveld (1977, 1978) concluded that the SIME-DIME NIT program resulted in a higher separation rate. These papers are summarized in their chapter in the SIME-DIME Final Report where they claim that "the negative income tax (NIT) plans tested in SIME/DIME dramatically increased the rate at which marriages dissolved among white and black couples" (Gronenveld,

[46] Note that mean happiness levels at baseline are high – 85% for females and 80% for males. In part, the high levels may reflect the use of a 4-point scale which may yield higher scores than the 5-point scale (which typically contains a neutral middle category).

[47] While church attendance may not be the best proxy for social activities today, there is good reason why it was the only question included along this dimension in the Mincome surveys. Attending church regularly was common in the early 1970s central Canada –the fraction of the population with no religious affiliation was less than 5%– and was an important part of a community's social activity (Canada's Changing Religious Landscape, Pew Research Report, June 2013). Statistics Canada's General Social Survey allows for some useful analysis for the 1985 to 2020 period. In 1985, 50-60% of 'Mincome-aged' adult cohorts (birth years from the 1920s to 1950) attended church at least once per month depending on the specific cohort. For comparison purposes, church attendance in 2020 for the children of 'Mincome' heads (birth years 1960s and 1970s) is only 20% (Statistics Canada, 2021).

Hannan and Tuma, 1983, p. 259). The Gronenveld et. al findings contributed importantly to the political opposition to adopting a NIT for two-parent families.[48] As stated by Cain and Wissoker (1990, p. 1236):

"Although SIME/DIME, like three other social experiments with income-maintenance plans sponsored by the (then) Department of Health, Education, and Welfare, was designed primarily to estimate labor-supply responses, the findings on marital breakups have had the biggest effect…Gilbert Steiner (1981, p. 110), who reviewed the testimony in congressional hearings on welfare reform, wrote that "the Seattle-Denver evidence has persuaded key politicians that a guaranteed-income plan at levels the leaders of the country think it can afford is incompatible with maximizing family stability in the affected population."

Subsequent reanalysis of the Grovenveld et al studies by Cain and Wissoker (1990) using a different set of methodological decisions questioned these conclusions. However, a paper by Keely (1987) also used SIME-DIME data—but a different empirical methodology from those above—and also found a sizeable increase in divorce rates in the treatment group. Thus, the issue remains unresolved.

All of the analysis on divorce (including all estimates in the Final Report) used the SIME-DIME pooled data. As documented in this paper, it is questionable that estimates from SIME can be viewed as unbiased. We re-assess the divorce evidence by estimating models for DIME alone as well as for the Mincome experiment.[49] Moreover, using Mincome data from the 'husband-wife module,' we explore satisfaction with the marriage, which could have implications for marital status over the longer run.

Referring to Table 9, in the pooled SIME-DIME data (for replication purposes), treatment status suggests an increase in the probability of divorce of 4 percentage points, or about a 16% increase (statistically significant at the 1% level).[50] However, the finding of a higher propensity to divorce is entirely driven by SIME—when the two experiments are analysed separately there

[48] For example, an official report on lessons learned from the Income Maintenance Experiments noted this finding: "In one experiment (Seattle-Denver), there were disturbingly higher family split-ups reported among some treatment groups than control groups." (Report to the Honorable Daniel P. Moynihan United States Senate. Income Maintenance Experiments: Need to Summarize Results and Communicate the Lessons Learned. US General Accounting Office, Comptroller General of the United States, 1981).

[49] There are problems with the marital status variable in New Jersey and thus we do not analyze that experiment.

[50] We define a 'divorce' as at least one marital separation over the post random assignment period. Note that we do not observe in SIME-DIME whether the split is a divorce or separation. Thus, our treatment effects should be interpreted as the effect of the NIT offer on the probability of a marital split occurring at some point. As noted in the literature (see text for key papers), a methodological point contested by authors was how to code cases where the couple get back together before the end of the experiment, which does occur in a non-trivial number of cases in the data (among other methodological points). Of course, it is also the case that such couples *subsequently split again*. Ultimately, given the three year time period (for the majority of observations) we cannot truly examine 'permanent' marital separation in these experiments.

is no statistically significant treatment effect in DIME (coefficient of 1.5 percentage points) whereas the point estimate suggests an ITT of 7 percentage points in SIME and is statistically significant at the 1% level. We also find no effect on the probability of divorce in Mincome (point estimate close to zero).

Results shown in Table 8 indicate that the NIT offer raised satisfaction with the marriage by about 7 percentage points for men, and 12 percentage points for women—relative to the control group. These point estimates imply percentage increases in marital satisfaction of 8% and 17% respectively (even greater for actual receipt of the NIT), increases that could offset any financial incentives that might exist for couples to separate.

Given the estimates discussed above, it seems plausible that the NIT offer could lead to positive longer-term implication of marriage stability. To shed some further light on this, we correlate our household-marital satisfaction variables—but measured *at the time of random assignment*—to the likelihood of marital dissolution (at some point over the 3-year experiment) in the final column of Table 9. Marital satisfaction and agreement on household duties at the baseline are both correlated—with sizeable impacts in terms of magnitude—with reduced marital separation rates over the duration of the experiment. We note for clarity that while the Mincome data is balanced (i.e., at the baseline the treatment and controls are equivalent in their household characteristics), the results in Table 9 simply indicate a correlation between agreement/satisfaction with various dimensions of the household/marriage and subsequent divorce. However, as documented above, households randomly assigned to the treatment group had greater levels of household task agreement and marital satisfaction as measured late in the experiment (approximately two years in). Thus, while our results provide no direct evidence, a suggestion of the experimental findings is that a NIT program may lead to positive long-run benefits to marital stability via raising marital satisfaction and preferred household allocation of duties.

6. Conclusions

The Negative Income Tax experiments carried out in North America in the late 1960s and early 1970s were the first large-scale field experiments conducted in economics. The evidence from these landmark studies is receiving renewed attention today given growing interest in a basic income to reduce poverty and inequality. While much policy discussion (and associated

pilot studies) revolves around a basic income being universal, careful assessments in the US and Canada question whether a BI that is not income-tested is fiscally realistic (Hoynes and Rothstein 2019; Green et.al. 2023; OECD 2017). As a consequence, interest in income-tested policies—including in particular a NIT—remains strong. For example, Barcelona, Spain and Seoul, South Korea have tested (or are testing) BI pilots with an NIT design. Nonetheless, the primary source of evidence on the probable impacts of introducing a basic income in developed countries remains the 1970s era NIT studies.

However, how reliable is this evidence? The economics profession has learned much about designing large-scale field experiments and analysing experimental data in the past 50 years. Our study employs contemporary empirical methods to re-assess the current consensus estimates, focusing on the behavior of male and female heads in two-parent families in the Income Maintenance Experiments carried out in New Jersey, Gary, Seattle and Denver and Winnipeg Manitoba.

A noteworthy feature of the current consensus regarding labor supply impacts for men and women in two-parent families is that, with the exception of women in Gary, all estimates on both the intensive and extensive margins are negative but only those for SIME-DIME (and Denver and Seattle separately) are statistically significant. Furthermore, some of the point estimates, especially those for men and women in Mincome, are near zero in magnitude (as well as statistically insignificant). Marinescu (2018) succinctly summarizes the international view on the North American NIT experiments: "The NIT experiments had a negative impact on employment among treated families, but this effect was generally not statistically significant." The inconsistent findings across experiments – both in the magnitudes of estimated responses and their statistical significance – "opens the way to alternative interpretations of the research findings" (Solow, 1986).

Our re-assessment demonstrates that the current consensus is incorrect. The labor supply estimates from the North American NIT experiments are actually remarkably robust and—other than men in New Jersey—quantitatively large. The ITTs range from -21 to -36% along the intensive margin and -17 to -32% along the extensive margin for women (-5 to -10% range on both margins for men).

While the original NIT research focused on impacts on labor supply and marital stability, most recent basic income research also investigates possible effects on various dimensions of

well-being. That said, one potentially important feature of a basic income that is often emphasized by BI advocates but less frequently incorporated into modern basic income trials (with some exceptions such as Vivalt et. al. 2024) is the notion that households could allocate time away from the labor market and towards other activities—in particular household production—that may yield important personal or social benefits. We examine hitherto untapped data from Mincome on subjective well-being as well as measures of household time allocation including household production, along with social activities. We find that the treatment group reports higher levels of happiness, marital satisfaction, agreement and satisfaction with household duties, in addition to increases in social activities (for women). Extrapolated to a modern labor supply setting, the findings suggest that an NIT income payment may allow households to re-allocate time in a manner that increases family well-being. These results may surprise some readers as one of the most concerning results from the US IMEs– especially in the SIME-DIME Final Report and subsequent Congressional testimony – was the negative impact of the NIT offer on marital stability. Subsequent analysis did not settle the issue. Our analysis concludes that this controversial result can be solely attributed to the Seattle site, where randomization appears to have failed. We find no evidence of the NIT offer on the likelihood of marital separation in Denver or Mincome with point estimates of nearly zero.

Labor market behavior has changed dramatically since the 1970s when these pioneering social experiments were carried out. Although this paper's main objective is to report new findings based on previously unexploited data as well as to 'set the record straight' about the key findings of these experiments, it is also worth speculating about the relevance of our empirical results for the ongoing BI debate. One key development has been the substantial rise in female labor force participation and the growing similarity of male and female labor force behavior. A closely related trend is the steady decline in the elasticities that influence female labor supply responsiveness – the elasticities of both hours worked and labor force participation with respect to changes in wage rates and income (e.g. Heim (2007) for the US and Dostie and Kromann (2013) for Canada). However, there are differences in responsiveness across the distribution of hours worked. Dostie and Kromann (2013) find that married women in the first decile of the hours of work distribution are much more responsive to wage changes than those in higher deciles. Thus, despite the substantial decline in elasticities for females as a whole, the families

that are most likely to be eligible for BI support may be the ones most likely to alter their work activity.

In addition, the availability of child care, including subsidized care for children in low-income families, has expanded substantially. This development may make the labor supply behavior of families with pre-school age children more similar to that of families with older children, perhaps lessening reductions in work activity that would otherwise occur. For this and for many other reasons, NIT-type experiments in today's labor markets are needed to supplement the evidence from the past.

References

Aaron, Henry J. 1975. Cautionary notes on the experiment. In Work incentives and income guarantees: The New Jersey negative income tax experiment, ed. Joseph A. Pechman and P. Michael Timpane. Washington, D.C.: The Brookings Institution.

Ardington, Cally, Anne Case and Victoria Hosegood. 2009. "Labor Supply Responses to Large Social Transfers: Longitudinal Evidence from South Africa" American Economic Journal: Applied Economics 1 (2009) 22-58.

Ashenfelter, Orley. 1978. "The Labor Supply Response of Wage Earners" in John L. Palmer and Joseph A. Pechman, eds. Welfare in Rural Areas: The North Carolina – Iowa Income Maintenance Experiment. Washington, D.C.: The Brookings Institution.

Ashenfelter, Orley and Mark W. Plant. 1990. Nonparametric Estimates of the Labor-Supply Effects of Negative Income Tax Programs. Journal of Labor Economics, 8: S396-S415.

Athey, Susan and Guido W. Imbens. 2017. The Econometrics of Randomized Experiments. In Handbook of Economic Field Experiments, Volume 1, ed. Esther Duflo and Abhijit Banergee . Amsterdam: Elsevier.

Banerjee, Abhijit V., Rema Hanna, Gabriel E. Kreindler and Benjamin A. Olken, 2017. "Debunking the Stereotype of the Lazy Welfare Recipient: Evidence from Cash Transfer Programs" The World Bank Research Observer, World Bank, 32(2), 155-184.

Banerjee, A., P. Neihaus and T. Suri. 2019. Universal Basic Income in the Developing World. Annual Review of Economics 11 961-985.

Banerjee, Abhijit, Michael Faye, Alan Krueger, Paul Niehaus and Tavneet Suri. 2020. Effects of a Universal Basic Income during the Pandemic. Working paper, December 8.

Bennett, S.E. 1986. "Do public program clients understand public program rules: The Manitoba Basic Annual Income Experiment." Canadian Public Policy/Analyse de Politiques, 29 (3): 462-68.

Bhargava, Saurabh and Dayanand Manoli. 2015. "Psychological Frictions and the Incomplete Take-Up of Social Benefits: Evidence from an IRS Field Experiment" American Economic Review 105 (11) 3489-3529.

Boadway, Robin, Katherine Cuff and Kourtney Koebel. 2018. Designing a Basic Income Guarantee for Canada. In Federalism and the Welfare State in A Multicultural World edited by E. Goodyear-Grant, R. Johnson, W. Kymlicka and J. Myles. Montreal and Kingston: McGill Queens University Press.

Burtless, Gary. 1987. The Work Response to a Guaranteed Income: A Survey of the Experimental Evidence. In Lessons from the Income Maintenance Experiments*, e*d. Alicia H. Mundell. Boston: Federal Reserve Bank of Boston Conference Series No. 30, pp. 22-52.

Cain, Glen G. and Douglas A. Wissoker (1990). A Reanalysis of Marital Stability in the Seattle-Denver Income-Maintenance Experiment. American Journal of Sociology. 95 (5): 1235-1269

Conlisk, John and Harold W. Watts. A Model for Optimizing Experimental Designs for Estimating Response Surfaces. Proceedings of the American Statistical Association (1969): 150-156.

Currie, Janet. 2006. "The Take-up of Social Benefits" in Public Policy and the Distribution of Income. Alan J. Auerbach, David Card and John M. Quigley (eds.). New York: Russell Sage Foundation, pp. 80-148.

Dostie, Benoit and L. Kromann. New Estimates of Labour Supply Elasticities for Married Women in Canada 1996-2005. Applied Economics 45 (31) 4355-68.

Fitzgerald, John, Peter Gottschalk and Robert Moffitt. 1998. An Analysis of Sample Attrition in Panel Data: The Michigan Panel Study of Income Dynamics. Journal of Human Resources 33: 300-344.

Forget, Evelyn. 2018. Basic Income for Canadians. Toronto: James Lorimer.

Garfinkel, Irwin. 1977. "Effects of welfare programs on experimental responses" chapter 12 in The New Jersey Income Maintenance Experiment, volume III edited by Harold W. Watts and Albert Rees. New York: Academic Press, pp. 279-302.

Green, David A., Jonathan Rhys Kesselman and Lindsay M. Tedds. 2020. Covering All the Basics: Reforms for a More Just Society. Final Report of the British Columbia Expert Panel on Basic Income. Victoria: Government of British Columbia.

Green, David A., Jonathan Rhys Kesselman, Daniel Perrin, Gillian Petit and Lindsay M. Tedds with France St-Hilaire. 2023. Basic Income and a Just Society: Policy Choices for Canada's Safety Net. Montreal: Institute for Research on Public Policy.

Greenberg, David and Harlan Halsey. 1983. Systematic Misreporting and Effects of Income Maintenance Experiments on Work Effort: Evidence from the Seattle-Denver Experiment. Journal of Labor Economics, 1: 380-407.

Greenberg, David, Robert Moffitt and John Friedmann. 1981. Underreporting and Experimental Effects on Work Effort: Evidence from the Gary Income Maintenance Experiment. Review of Economics and Statistics, 63: 581-89.

Gronenveld, Lyle P., Michael T. Hannan and Nancy B. Tuma. 1983. "Marital Stability." Pp. 257-387 in Final Report of the Seattle-Denver Income Maintenance Experiment, vol. 1. Design and Results. Menlo Park, Calif.: SRI International.

Haagh, Louise. 2019. The Case for Universal Basic Income. London: Polity Press.

Handa, Sudhanshu, Kuisa Natali, David Seidenfeld, Gelson Tembo and Benjamin Davis. 2018. Can unconditional cash transfers raise long-term living standards? Evidence from Zambia. Journal of Development Economics 133, 42-65.

Hanna, Rema and Benjamin A. Olken. 2018. Universal Basic Incomes versus Targeted Transfers: Anti-Poverty Programs in Developing Countries. Journal of Economic Perspectives 32 (4): 201-226.

Hannan, Michael T., Nancy B. Tuma, and Lyle P. Groeneveld. 1977. "Income and Marital Events: Evidence from an Income-Maintenance Experiment." American Journal of Sociology 82: 1186-1211.

Hannan, Michael T., Nancy B. Tuma, and Lyle P. Groeneveld. 1978. "Income and Independence Effects on Marital Dissolution: Results from the Seattle and Denver Income-Maintenance Experiments." American Journal of Sociology 84: 611-33.

Haushofer, Johannes and Jeremy Shapiro. 2016. "The Short-Term Impact of Unconditional Cash Transfers to the Poor: Experimental Evidence from Kenya" Quarterly Journal of Economics 1973-2042.

Haushofer, Johannes, Matthieu Chemin, Chaning Jang and Justin Abraham. 2019. Economic and Psychological Effects of Health Insurance and Cash Transfers: Evidence from a Randomized Experiment in Kenya. Working paper, November 11.

Heim, Bradley T. 2007. "The Incredible Shrinking Elasticities: Married Female Labor Supply, 1978-2002" Journal of Human Resources 42 (4) 881-918.

Hernanz, V., F. Malherbet and Pellizzari. 2004. Take-Up of Welfare Benefits in OECD Countries: A Review of the Evidence. OECD Social, Employment and Migration Working Papers #17, March.

Hoynes, Hilary and Jesse Rothstein. 2019. Universal Basic Income in the U.S. and Advanced Countries. Annual Review of Economics, 11:929-958.

Hum, Derek and Wayne Simpson. 1991. Income Maintenance, Work Effort and the Canadian Mincome Experiment. Ottawa: Economic Council of Canada.

Hum, Derek and Wayne Simpson. 1993. Economic Response to a Guaranteed Annual Income: Experience from Canada and the United States. Journal of Labor Economics, 11: S263-S296.

Jacob, Brian A., Pilkauskas, Natasha V., Elizabeth Rhodes, Katherine Richard and H. Luke Shaefer. 2022. The COVID Cash Transfer Study: The Impacts of an Unconditional Cash Transfer on the Wellbeing of Low-Income Families. May.

Jaroszewicz, Ania, Oliver P. Hauser, Jon M. Jachimowicz, and Julian Jamison. 2024. How Effective is (More) Money? Randomizing Unconditional Cash Transfer Amounts in the US. SSRN Working Paper 4154000. July.

Keeley, Michael C. 1987. The Effects of Experimental Negative Income Tax Programs on Marital Dissolution: Evidence from the Seattle and Denver Income Maintenance Experiments. International Economic Review 28 (1): 241-257.

Keeley, Michael C. and Philip K. Robins. 1980. Experimental Design, the Conlisk-Watts Assignment Model and the Proper Estimation of Behavioral Response. Journal of Human Resources, 13: 3-36.

Kershaw, David and Jerilyn Fair. 1976. The New Jersey Income Maintenance Experiment: Volume 1 Operations, Surveys and Administration. New York: Academic Press.

Levine, Robert A. 1975. How and Why the Experiment Came About. In Work Incentives and Income Guarantees: The New Jersey negative income tax experiment, ed. Joseph A. Pechman, and P. Michael Timpane. Washington, D.C.: The Brookings Institution, 15-24.

Londono-Velez, Juliana and Pablo Querubin. 2022. The Impact of Emergency Cash Assistance in a Pandemic: Experimental Evidence from Columbia. Review of Economics and Statistics 104 (1) January 157-165.

Lowrey, Annie. 2018. Give People Money: How A Universal Basic Income Would End Poverty, Revolutionize Work, and Remake the World. Crown Publishing.

Marinescu, Ioana. 2018. No Strings Attached: The Behavioral Effects of U.S. Unconditional Cash Transfer Programs. NBER Working Paper 24337, February.

McGuire, Joel, Caspar Kaiser and Anders M. Bach-Mortensen. 2022. A systematic review and meta-analysis of the impact of cash transfers on subjective well-being and mental health in low- and middle-income countries. Nature Human Behaviour 6 (March) 359-370.

Moffitt, Robert A. 1979.The Labor Supply Response in the Gary Experiment. Journal of Human Resources.14 (4) 477-487.

Moffitt, Robert A. 2003. The Negative Income Tax and the Evolution of U.S. Welfare Policy. Journal of Economic Perspectives, 17: 119-140.

Murray Charles. 2016. In Our Hands: A Plan to Replace the Welfare State. Washington, D.C: AEI Press.

New York Times. 1971. “Layoffs Spread at Steel Mills” August 7.

OECD. 2017. Basic Income as a Policy Option: Can It Add Up?. Paris: OECD, May.

Pechman, Joseph A. and P. Michael Timpane. 1975. Introduction and summary. In Work incentives and income guarantees: The New Jersey negative income tax experiment, ed. Joseph A. Pechman, and P. Michael Timpane. Washington, D.C.: The Brookings Institution.

Pilkauskas, N.V., Jacob, B.A., Rhodes, E., Richard, K. & Shaefer, H. L. (2023). The COVID cash transfer study: the impacts of a one-time unconditional cash transfer on the wellbeing of families receiving SNAP in twelve states. Journal of Policy Analysis and Management 42 (3) 771-795.

Price, David J. and Jae Song. 2018. The Long-Term Effects of Cash Assistance. Princeton University, Industrial Relations Section, Working Paper #621 (Revise and Resubmit: American Economic Journal: Applied Economics).

Rainey, R.B. Jr. et.al. 1973. Seattle’s Adaption to Recession. Santa Monica: The Rand Corporation, September.

Riddell, Chris and W. Craig Riddell. 2024. Welfare versus Work under a Negative Income tax: Evidence from the Gary, Seattle, Denver and Manitoba Income Maintenance Experiments. Journal of Labor Economics 42 (2) April.

Riutort, Sebastia, Bru Lain and Albert Julia. 2023. Basic Income at Municipal Level: Insights from the Barcelona B-MINCOME Pilot. Basic Income Studies 18 (1) 1-30.

Robins, Philip. 1985. A Comparison of the Labor Supply Findings of the Four Negative Income Tax Experiments. Journal of Human Resources, 20: 567-82.

Robins, Philip K. and Richard W. West. 1980. Program Participation and Labor Supply Response. Journal of Human Resources, 15: 499-523.

Schirle, Tammy. 2008. Why Have the Labour Force Participation Rates of Older Men Increased Since the Mid-1990s? Journal of Labor Economics 26 (4) 549-594.

Simpson, Wayne, Greg Mason and Ryan Godwin. 2017. The Manitoba Basic Annual Income Experiment: Lessons Learned 40 Years Later. Canadian Public Policy, 43: 85-104.

Skidmore, Felicity. 1976. Operations and Administration. In The New Jersey Income-Maintenance Experiment Volume 1, D. Kershaw and J. Fair eds. Academic Press.

Solow, Robert M. 1987. “An Economist’s View of the Income Maintenance Experiments” In Lessons from the Income Maintenance Experiments, *e*d. Alicia H. Mundell. Boston: Federal Reserve Bank of Boston Conference Series No. 30, pp. 218-222.

Spiegelman, R. and K. Yaeger. 1980. Overview. Journal of Human Resources 15 463-479.

SRI International. 1983. Final Report of the Seattle-Denver Income Maintenance Experiment. Menlo Park, Ca.: SRI International.

Van Parijs, Philippe and Yannick Vanderborght. 2017. Basic Income: A Radical Proposal for a Free Society and a Sane Economy. Cambridge, Mass.: Harvard University Press.

Verho, Juoko, Kari Hamalainen and Ohto Kanninen. 2022. Removing Welfare Traps: Employment Responses in the Finnish Basic Income Experiment. American Economic Journal: Economic Policy. 14(1) 501-522.

Vivalt, Eva, Elizabeth Rhodes, Alexandar W. Bartik, David E. Broockman and Sarah Miller. 2024. The Employment Effects of a Guaranteed Income: Experimental Evidence from Two U.S. States. NBER Working Paper 32719. July.

Watts, Harold W. and Albert Rees. 1976. The New Jersey Income Maintenance Experiment: Volume II Labor-Supply Responses. New York: Academic Press.

Widerquist, Karl. 2005. A failure to communicate: what (if anything) can we learn from the negative income tax experiments? The Journal of Socio-Economics, 34: 49-81.

World Bank (2020). Exploring Universal Basic Income: A Guide to Navigating Concepts, Evidence, and Practices, edited by Ugo Gentilini, Margaret Grosh, Jamele Rigolini, and Ruslan Yemtsov. Washington, DC: World Bank.

Yang Andrew. 2018. The War on Normal People: The Truth About America’s Disappearing Jobs and Why Universal Basic Income Is Our Future. Hachette Books.

Table 1
Balancing Tests for Mincome

	Women						Men					
	Received Welfare	Employed	Hours Worked	Years of Education	Age	Youngest Child <6	Received Welfare	Employed	Hours Worked	Years of Education	Age	Youngest Child <6
Treatment Group	-.033 (.024)	-.021 (.039)	-8.29 (36.02)	-.025 (.217)	-1.21 (.753)	.045 (.038)	-.030 (.024)	.007 (.024)	70.11 (53.76)	.090 (.221)	-.932 (.853)	.039 (.038)
Mean Dependent Variable	.116	.374	258.8	9.65	30.67	.659	.114	.877	1022.37	9.73	33.17	.658
Number of Individuals	821	790	790	817	821	790	821	788	788	818	821	789

NOTES—Standard errors are in parentheses. The baseline data for Mincome is an annual cross-section. All regressions include controls for experimental stratification ('income cell' which is family size-adjusted pre-random assignment income). Estimation is by OLS. Statistical significance indicated as follows: * for 10% level, ** for 5% level, and *** for 1% level.

Table 2
Balancing Tests for DIME

	Women						Men					
	Received Welfare	Employed	Hours Worked	Earnings	Age	Youngest Child <6	Received Welfare	Employed	Hours Worked	Earnings	Age	Youngest Child <6
Treatment Group	.025* (.015)	.009 (.019)	2.33 (2.94)	2.72 (6.60)	-.082 (.491)	.012 (.025)	.022 (.015)	.009 (.014)	-3.92 (3.19)	-5.67 (12.05)	.164 (.500)	.006 (.025)
Mean Dependent Variable	.139	.289	40.86	83.05	29.36	.638	.137	.842	150.88	470.31	32.14	.638
Number of Individuals	1459	1459	1459	1458	1459	1459	1503	1503	1503	1499	1503	1503
Sample Size	21829	21814	21814	21684	21829	21828	22519	22518	22518	22077	22519	22518

NOTES—Standard errors are in parentheses, and are clustered on the individual due to multiple months of pre-random assignment data. Denver is a monthly dataset. All regressions include controls for experimental stratification (interactions of family size-adjusted pre-random assignment income, race, and program length). Estimation is by OLS. Statistical significance indicated as follows: * for 10% level, ** for 5% level, and *** for 1% level.

Table 3
Balancing Tests for New Jersey

	Women						Men					
	Received Welfare	Employed	Hours Worked	White	Age	Youngest Child <6	Received Welfare	Employed	Hours Worked	White	Age	Youngest Child <6
Treatment Group	-.006 (.023)	.030 (.021)	9.31 (8.73)	.011 (.017)	.107 (.505)	.043 (.028)	-.000 (.023)	-.022 (.019)	-20.81* (12.44)	.012 (.017)	.395 (.547)	.032 (.029)
Mean Dependent Variable	.184	.150	59.40	.329	33.76	.470	.175	.874	438.88	.333	37.06	.473
Number of Individuals	1195	1297	1297	1306	1306	1306	1148	1248	1248	1254	1254	1253

NOTES—Standard errors are in parentheses. New Jersey is a quarterly data set, and there is only one quarter of pre-random assignment data; hence this is a cross-section. All regressions include controls for experimental stratification (interactions of family size-adjusted pre-random assignment income and site). Estimation is by OLS. Statistical significance indicated as follows: * for 10% level, ** for 5% level, and *** for 1% level.

Table 4
Balancing Tests for Mincome: Marital Satisfaction and Household Time Allocation

	Women			Men		
	Agree on who does housework	Satisfied with partner's help around house	Happy with marriage	Agree on who does housework	Satisfied with partner's help around house	Happy with marriage
Treatment Group	.012 (.042)	-.003 (.039)	.024 (.029)	-.031 (.042)	-.010 (.030)	-.040 (.027)
Mean Dep. Variable (Controls)	.498	.690	.866	.575	.889	.883
Number of Individuals	690	694	691	688	695	702

NOTES—Standard errors are in parentheses. Estimation is based on the Enrolment Survey. All regressions include controls for experimental stratification (income cell). Estimation is by OLS. Statistical significance indicated as follows: * for 10% level, ** for 5% level, and *** for 1% level.

Table 5
Tests for Attrition Bias: Mincome

	Women				
	Employed	Hours Worked	Agree on who does housework	Satisfied with partner's help around house	Happy with marriage
Attritor	.046 (.040)	19.39 (36.39)	.019 (.046)	-.053 (.043)	-.027 (.030)
Number of observations	794	794	644	638	645
	Men				
	Employed	Hours Worked	Agree on who does housework	Satisfied with partner's help around house	Happy with marriage
Attritor	-.043* (.025)	70.09 (52.99)	.002 (.044)	-.034 (.033)	-.038 (.028)
Number of observations	793	793	636	639	641

NOTES—Standard errors are in parentheses. Estimation is by OLS. All regressions include demographic controls. Labor market outcomes are measured over the pre-experimental period, and attritor=1 if the individual would drop-out at any point over the experimental period (=0 if the individual remains in the experiment until its conclusion). Statistical significance denoted by *** for the 1% level, ** for the 5% level and * for the 10% level.

Table 6
Tests for Attrition Bias: DIME and New Jersey

	DIME			
	Women		Men	
	Employed	Hours Worked	Employed	Hours Worked
Attritor	-.010 (.020)	-2.53 (3.17)	-.016 (.015)	-3.76 (3.41)
Number of Individuals	1459	1459	1503	1503
Sample Size	21814	21814	22518	22518
	New Jersey			
	Women		Men	
	Employed	Hours Worked	Employed	Hours Worked
Attritor	.012 (.026)	16.30 (11.82)	-.021 (.022)	-32.90** (14.00)
Number of Individuals	1297	1297	1247	1247
Sample Size	1297	1297	1247	1247

NOTES—Standard errors are in parentheses, and are clustered on the individual. Estimation is by OLS. All regressions include demographic controls. Labor market outcomes are measured over the pre-experimental period, and attritor=1 if the individual would drop-out at any point over the experimental period (=0 if the individual remains in the experiment until its conclusion). Statistical significance denoted by *** for the 1% level, ** for the 5% level and * for the 10% level.

Table 7
Estimated Treatment Effects: Labor Supply

	Mincome, Women		Mincome, Men		DIME, Women		DIME, Men		New Jersey, Women		New Jersey, Men	
	Hours Worked	Employed	Hours Worked	Employed	Hours Worked	Employed	Hours Worked	Employed	Hours Worked	Employed	Hours Worked	Employed
Treatment Group	-47.65*** (16.73)	-.070** (.032)	-40.76** (19.04)	-.046* (.025)	-8.40*** (2.81)	-.062*** (.017)	-14.41*** (3.13)	-.046*** (.013)	-21.59*** (6.21)	-.048*** (.014)	-10.77 (9.22)	-.007 (.015)
Mean Dep. Variable (Controls)	180.62	.414	492.70	.84	40.86	.288	150.88	.842	59.4	.149	438.9	.874
ITT in %	-26%	-17%	-8%	-5%	-21%	-22%	-10%	-6%	-36%	-32%	-2%	0%
Coefficient in annual hours	144	-	123	-	101	-	173	-	86	-	35	-
Number of Individuals	690	690	697	697	1424	1424	1473	1473	1267	1267	1208	1208
Sample Size	4735	4735	4777	4777	55836	55836	55641	55641	13309	13309	11995	11995

NOTES—Standard errors are in parentheses, and are clustered on the individual. Statistical significance denoted by *** for the 1% level, ** for the 5% level and * for the 10% level. All regressions include fixed effects for the experimental stratification (income cell*site for New Jersey, income cell for Mincome, and income cell*race*program length for DIME) and survey/time, in addition to demographic controls and pre-random assignment employment. Estimation is by OLS.

Table 8
Estimated Treatment Effects: Happiness and Household Time Allocation

	Women					Men				
	Happy	Agree on who does housework	Satisfied with partner's help around house	Happy with marriage	Church visits	Happy	Agree on who does housework	Satisfied with partner's help around house	Happy with marriage	Church visits
Treatment Group	.046** (.021)	.121** (.051)	.155*** (.049)	.124*** (.041)	1.01*** (.263)	.052** (.024)	.041 (.050)	.067* (.037)	.066* (.035)	-.078 (.286)
Mean Dep. Variable (Controls)	.847	.443	.574	.754	4.80	.800	.514	.834	.853	4.60
Number of Individuals	766	479	496	496	374	679	482	483	496	345
Sample Size	1870	479	496	496	766	1603	482	483	496	704

NOTES—Standard errors are in parentheses, and are clustered on the individual where appropriate. Statistical significance denoted by *** for the 1% level, ** for the 5% level and * for the 10% level. The online appendix (section 3) provides definitions of all variables. All regressions include fixed effects for the experimental stratification (income cell) and survey/time, in addition to demographic controls and pre-random assignment employment. Estimation is by OLS.

Table 9
Estimated Treatment Effects: Probability of a Marital Split

	SIME	DIME	SIME-DIME	Mincome	Mincome (Extended Controls)
Treatment Group	.068*** (.026)	.015 (.025)	.038** (.018)	-.016 (.026)	.010 (.024)
Agree on who does housework	-	-	-	-	-.074*** (.025)
Satisfied with husband's help around house	-	-	-	-	-.017 (.028)
Happy with marriage	-	-	-	-	-.072** (.035)
Mean Dependent Variable (Control Group)	.212	.274	.244	.246	.232
Sample Size (Number of Households)	1121	1461	2582	821	722

NOTES—Standard errors are in parentheses. Statistical significance denoted by *** for the 1% level, ** for the 5% level and * for the 10% level. All regressions include fixed effects for the experimental stratification. Additional Mincome controls in column five are measured at random assignment. The sample here is cross-sectional where the outcome variable =1 if a household experiences a marital split (divorce or separation) at any time over the experimental period (=0 if not).

Table A1
Balancing Tests: Gary

	Women						Men					
	Received Welfare	Employed	Hours Worked	Earnings	Age	Youngest Child <6	Received Welfare	Employed	Hours Worked	Earnings	Age	Youngest Child <6
Treatment Group	-.058*** (.023)	.015 (.048)	-.170 (1.91)	-27.67 (62.17)	.149 (1.02)	-.019 (.051)	.009 (.009)	.059** (.024)	1.90* (1.09)	-24.47 (62.50)	.866 (1.47)	-.007 (.053)
Mean Dependent Variable	.069	.219	7.75	638.8	39.13	.357	.014	.890	35.50	658.7	42.29	.363
Number of Individuals	540	564	564	582	582	582	550	541	541	557	557	557
Sample Size	2195	2275	2275	2349	2349	2349	2263	2193	2193	2279	2279	2279

NOTES—Standard errors are in parentheses, and are clustered on the individuals. All regressions include controls for experimental stratification (interactions of family size-adjusted pre-random assignment income and locations). Estimation is by OLS. Statistical significance indicates as follows: * for 10% level, ** for 5% level, and *** for 1% level.

Table A2
Balancing Tests: SIME

	Women						Men					
	Received Welfare	Employed	Hours Worked	Earnings	Age	Youngest Child <6	Received Welfare	Employed	Hours Worked	Earnings	Age	Youngest Child <6
Treatment Group	.045*** (.012)	-.025 (.023)	-6.06* (3.58)	-47.11*** (8.21)	-.605 (.590)	-.011 (.028)	.044*** (.012)	-.013 (.019)	-2.60 (.2.98)	-159.05*** (14.23)	.158 (.627)	-.010 (.028)
Mean Dependent Variable	.091	.300	42.84	85.81	32.77	.583	.090	.731	123.92	392.43	35.56	.583
Number of Individuals	1157	1157	1157	1156	1157	1157	1156	1156	1156	1156	1156	1156
Sample Size	19073	19073	19073	18944	19073	19073	19062	19062	19062	19062	19062	19062

NOTES—Standard errors are in parentheses, and are clustered on the individual for regressions based on multiple months pre-random assignment. All regressions include controls for experimental stratification (interactions of family size-adjusted pre-random assignment income and race). Estimation is by OLS. Statistical significance indicated as follows: * for 10% level, ** for 5% level, and *** for 1% level.

ONLINE APPENDIX

1. Consensus Estimates of Treatment Effects for Two-Parent Families[1]

	Men		Women	
	Annual hours	Employment rate	Annual hours	Employment rate
		NEW JERSEY		
Robins (1985)	-34.2	-.01	-55.7	-.04
Burtless (1987)	-21 (1.2%)		-56 (24.6%)	
		GARY		
Robins (1985)	-35.4	-0.01	-57.6	-0.03
	(65.1)	(.03)	(62.7)	(.04)
Burtless (1987)[2]	-114 (6.5%)	N/R	+14 (5.0%)	N/R
		SIME – DIME		
Robins (1985)	-112.8***	-0.04***	-141.2***	-0.08***
	(30.1)	(.01)	(34.5)	(.02)
Burtless (1987)	-133 (7.1%)	N/R	-101 (14.2%)	N/R
SRI International (1983)	-133.1***	-0.05**	-101.4***	-0.11**
	(37.4)	(.01)	(35.8)	(.02)
		MINCOME		
Hum and Simpson (1993)[3]	-17 (1%)		-15 (3%)	
		DIME		
SRI International (1983)	-149.6***	N/R	-103.2*	N/R
	(54.6)		(51.6)	
SRI International (1983)		SIME		
	-123.0**	N/R	-100.2*	N/R
	(50.8)		(49.3)	

[1] Robins and Burtless present treatment effects that are weighted averages calculated from selected samples from the Final Reports, and thus are not identical. Standard errors are in parentheses below the estimated coefficients. Standard errors and statistical significance were not reported for New Jersey. Percentage changes relative to baseline are in parentheses beside estimated coefficients. ***, ** and * indicate statistical significance at 1%, 5% and 10% levels.

[2] Burtless (1987) and Hum and Simpson (1993) do not report employment rate estimates (denoted N/R) or standard errors.

[3] Estimates for men include single men (21% of all males in sample).

2. Mincome: Well-being, Time Use and Related Survey Questions

Marital Satisfaction and Time Allocation:

Helping with work around the house.

Always disagree	1
Almost always disagree	2
Occasionally disagree	3
Almost always agree	4
Always agree	5

As someone who is helpful around the home.

Very dissatisfied	1
Dissatisfied	2
Neither satisfied nor dissatisfied	3
Satisfied	4
Very satisfied	5

In general, how happy would you say you are with your marriage?

Very unhappy	1
Unhappy	2
Neither happy nor unhappy	3
Fairly happy	4
Very happy	5

Happiness:

Taken altogether, how would you say things are these days…would you say you were very happy, fairly happy, not too happy, or not happy at all?

Very happy	1
Fairly happy	2
Not too happy	3
Not happy at all	4

Church Attendance:

How frequently do you go to worship services?

Never	1
Several times a year	2
Once a month	3
Two-three times a month	4
Every week	5
More than once a week	6

Data Availability by Survey:

Baseline Survey	
Enrollment Survey	Husband-Wife Module
1st Periodic	
2nd Periodic	Happiness/Church Questions
3rd Periodic	
4th Periodic	
5th Periodic	Husband-Wife Module
6th Periodic	Happiness/Church Questions
7th Periodic	
8th Periodic	
9th Periodic	Happiness/Church Questions

3. Replication of Hum and Simpson (1991; 1993)

a) Sample sizes:

We have tried to replicate the findings of Hum and Simpson across all family types, and have been unable to do so. Indeed, we cannot even replicate their sample. Here, we focus on the sample sizes for the Two-Heads, Winnipeg experiment (also referred to in some Mincome documentation as "Double-Headed").[4, 5] All public use files with post random assignment data including the data Hum and Simpon report to utilize (1991, see page 58) and the official Mincome documentation (see, for example, page 1 of the Minc4 record layout: "Mincome Longitudinal Labour Data File"[6]) state that the public use file is for "In-Tact" families. From the Minc4 record layout "In-Tact" has the same definition as noted in the paper with respect to New Jersey: "households in which there was no head split (divorce/separation) or head join (marriage including common law).[7] For clarity, we note that the baseline survey, referred to as the Minc1 file ("MINCOME Baseline Summary File")[8], contains data on almost all of the non "In-Tact" households, but post random assignment labor market (and other) data in Minc4 was only digitized for the "In-Tact" households.

Table A.3.1 below shows cell counts for the Two Head, Winnipeg experiment for different sample constructions. We emphasize that Minc4 should be the starting sample for estimation of treatment effects; that is, the columns in italics should be

[4] The Minc4 record layout defines "double-headed" as: households with both a male and female head present with or without any children (we note that, in Mincome, there are very few two-head households with no children).
[5] As noted in the text: the Rural experiment has never been digitized (or even accessed); the Dauphin data is non-experimental and has never been digitized (for the survey data; Minc2 Dauphin has been utilized); and the Winnipeg-Supplemental experiment has never been digitized or even accessed (according to the staff at the Archives of Canada) until this paper.
[6] Institute for Social and Economic Research, 1983 (https://gregorymason.ca/mincome/).
[7] As noted below in Section 4, this is incorrect as there are a relatively large number of cases in the non in-tact data that are other situations of a household head 'breaking' from the original enrolled unit including deaths, incarcerations, and in particular various idiosyncratic household composition cases (such as where an adult child is the original head, departs and then that individual's parents become the new heads post random assignment).
[8] Institute for Social and Economic Research, 1983 (https://gregorymason.ca/mincome/).

the same. For completeness we also document Minc1 here, and remind readers that our Mincome sample in the paper includes these non "In-Tact" households. The total sample size listed in both the Minc1 and Minc4 record layouts matches ours (and the public use data) with a total "Double-Headed", Winnipeg experiment baseline sample of n=794, and a total sample of n=701 for Minc4. Focusing on the Minc4 data, inspection of the treatment variable in the microdata shows that of these 701 households, 195 were never enrolled in the experiment (and thus are missing all post random assignment data, see below for further details), 181 households were randomly assigned to the control group, and 325 household were randomly assigned to the treatment group (with their specific plan listed in the Table). Hum and Simpson, however, report 302 treatment group households and 348 control group households. The reasons for the smaller number of treatments and much larger number of controls are unclear although it is noteworthy that the treatment group counts are close, and have a comparable distribution by treatment plan.[9] As noted in the main text, one consequence of the Conlisk-Watts assignment model as implemented in all four U.S. NITs is a smaller number of controls than treatments, a feature that also holds in our sample and the Mincome documentation. The fact that the number of control families in their sample exceeds treatments raises serious doubts about the validity of their sample.

One feature of Table A.3.1 that stands out is the greater comparability of the Hum and Simpson sample (N=650) to the full Minc4 data (N=701) vs. the enrolled sample (N=506). One possibility, that we discuss in more detail below in part b), is that the non-enrolled households were included in the Hum and Simpson analysis and coded as members of the control group. We document the treatment status counts under this scenario in Table A.3.1, and the control group counts are much more similar. Given that the treatment observations are very close (although it is perplexing why those differ at all), and that the Hum and

[9] One potential explanation of the differences in treatment group counts is that Hum and Simpson dropped (as we do in our sample selection criteria subsequently) those households whose income cell stratification is missing (at least missing in the current public use file). We cannot include these observations (unless we predict their income cell) as we do not observe their stratification. Thus, we also tabulate the treatment counts excluding these cases, but the treatment group counts still do not align with Hum and Simpson.

Simpson control group sample size must be overstated for reasons discussed above, we can find no other plausible scenario (aside from the public use data simple being erroneous, see below) for the control group sample size being so large.[10]

b) Hours worked errors in the public use file:

Related to above, in attempting to replicate Hum and Simpson we also discovered a key error in the hours worked variable (i.e., in the public use file). Specifically, hours worked that are missing—including, in particular, those households who were screened but never enrolled—were not coded as -9 (as the official Mincome documentation indicates was supposed to be the case), but rather are coded as a 0, i.e., the same number as enrolled individuals who did not work (for both men and women). This was never noted in the Hum and Simpson work (nor to the best of our knowledge anywhere else in the Mincome literature such as Simpson, Mason and Godwin (2017)). Given that a classic econometric model for estimating treatment effects specifies a post-random assignment outcome (such as hours worked) as a function of a treatment dummy (and, in this case, experimental stratifications) and —to reduce residual variance—pre-random assignment characteristics, the result of this error is that the non-enrolled (i.e., who never participated in the experiment) could be included in the analysis and count as zero hours worked if coded as part of the control group.[11] Such an error would substantially reduce the mean hours worked and mean employment rate for the control group, thus mechanically biasing treatment effects upwards (and would of course also increase the sample size of controls). As noted in the text, Hum and Simpson's estimates are negative in sign but close to zero (and not statistically different from zero).

[10] Moreover, the Mincome documentation (the Minc4 record layout specifically), unfortunately, lacked clarity on disclosing sample sizes as it somewhat misleadingly classifies non-enrolled households as "non-completers" (when these households were never actually randomly assigned).

[11] Coding the non-enrolled as part of the control group is plausible given the treatment plan classifications. We note that these households—i.e., screened but never enrolled—also appear in the SIME-DIME monthly principal person file, but in that case all of the post random assignment data is labelled with a missing value indicator (e.g., 9999).

Table A.3.1
Mincome Cell Counts across Samples: Two Head Households, Winnipeg Experiment

NIT Plan	(1) 'Minc1' (Includes non "In-Tact" Households)		(2) 'Minc4' ("In-Tact" Households)		(3) Hum and Simpson (1991, page 58 Table 7-1) —"Double-headed", "In-Tact"
	Drop -1	Include -1 (classify as Control Group)	Drop -1	Include -1 (classify as Control Group)	
G=$3800; t=0.35	39	39	*32*	32	*29*
G=$4800; t=0.35	45	45	*40*	40	*39*
G=$3800; t=0.50	55	55	*47*	47	*41*
G=$4800; t=0.50	72	72	*61*	61	*59*
G=$5800; t=0.50	43	43	*38*	38	*35*
G=$3800; t=0.75	50	50	*44*	44	*40*
G=$4800; t=0.75	30	30	*26*	26	*24*
G=$5800; t=0.75	46	46	*37*	37	*35*
Total Treatments	378	378	*325*	325	*302*
Control Group	218	414	*181*	376	*348*
Sample Size	598	794	*506*	701	*650*

NOTES: Minc1 counts are based on the stand-alone Baseline Survey (known as "Minc1" in the public use files), which (along with the Enrollment Survey) we use to estimate pre random assignment characteristics. Minc4 is the source of post random assignment labor market data but was only digitized for "In-Tact" households. For the columns 'Drop -1' we exclude those households who were deemed eligible for Mincome following the screening survey and subsequently were administered the Baseline Survey, but ultimately not enrolled in the experiment (i.e., were never randomly assigned). For the columns 'Include -1' we include the non-enrolled as members of the control group.

4. Audit/Data Compilation of Winnipeg Experiment

Table A.4.1
Cell Counts for Complete Winnipeg Experiment: Two Heads Sample

	Treatment Group	Control Group	Total Observations
Original Winnipeg Experiment, "In-Tact" Households (Minc4 public use file)	325	181	506
Original Winnipeg Experiment, non- "In-Tact" Households (never digitized)	62	33	95
Supplemental Winnipeg Experiment (never digitized)	151	69	220
Total Households Winnipeg Experiment for Two Heads	579	274	821

Table A.4.2
Audit of Original Winnipeg Experiment: Archives of Canada vs Public Use Microdata File Summary, All Households

(1) Total count in Minc1	(2) Total count Minc1 missing from Minc4 (non-"In-Tact")	(3) Total count non-"In-Tact" family IDs accounted for at Archives	(4) Total count Minc4	(5) Total count Minc4 never enrolled	(6) Total count in Minc4 enrolled	(7) Total count enrolled Minc4 accounted for at Archives	(8) Total count Archives of Canada family IDs missing from public use file	(10) Total count of available enrolled data public use + archives ((3) + (6) + (8))
1437	147	146	1290	369	921	913	6	1073

5. New Jersey Experiment: Additional Implementation Details and Robustness Checks

a) Aaron (1975) and Pechman and Timpane (1975) discuss concerns about the implications of the policy change for the interpretation of experimental findings. These fall into three categories – internal validity, external validity and take-up (although this terminology was not used at the time). Internal validity could be affected if control group males reduced their work activity relative to what would have occurred in the absence of the new policy. This 'substitution' of an alternative form of income support could reduce the size of the treatment effect relative to what it would have been without the policy change. External validity could be affected if large numbers of treatment and/or control families choose welfare over the NIT program, making some experimental cells too small to provide reliable estimates. This latter possibility also affects take-up. Garfinkel (1977) provides a detailed analysis of the effects of the new welfare policy on the experiment, with particular attention to internal validity. He concludes that the labor supply impacts for the eight experimental plans would not have been very different from those observed in the experiment in the absence of an AFDC-UP program in New Jersey and Pennsylvania. However, there may be some concern regarding external validity as many families originally enrolled in the least generous NIT plans (those with a 50% [of the poverty line] guarantee and 50% tax rate or 75% guarantee and 70% tax rate) chose welfare over the NIT. Since the selection model implies that these plans were more likely to be populated by lower income families, the experimental sample may be less representative of the low-income population.

Table A.5.1
ITT Estimates for New Jersey: Estimates across Experimental Site

	Excluding Trenton, Women		Excluding Trenton, Men		Excluding Trenton + Pennsylvania, Women		Excluding Trenton + Pennsylvania, Men		Pennsylvania, Women		Pennsylvania, Men	
	Hours Worked	Employed	Hours Worked	Employed	Hours Worked	Employed	Hours Worked	Employed	Hours Worked	Employed	Hours Worked	Employed
Treatment Group	-20.74*** (6.54)	-.049*** (.015)	-8.23 (9.70)	-.001 (.015)	-23.40*** (7.95)	-.055*** (.017)	7.29 (11.13)	.021 (.019)	-17.57* (10.88)	-.048** (.024)	-41.1** (19.40)	-.051* (.029)
Mean Dep. Variable (Controls)	53.6	.139	440.7	.882	53.6	.133	435.2	.873	53.6	.151	455.2	.901
ITT in %	-39%	-35%	-2%	0%	-44%	-41%	2%	2%	-33%	-32%	-9%	-6%
Number of Individuals	1121	1121	1073	1073	806	806	771	771	315	315	302	302
Sample Size	11886	11886	10739	10739	8251	8251	7366	7366	3636	3635	3374	3374

NOTES—Standard errors are in parentheses, and are clustered on the individual. Statistical significance denoted by *** for the 1% level, ** for the 5% level and * for the 10% level. All regressions include fixed effects for the experimental stratification (income cell*site) and survey/time, in addition to demographic controls and pre-random assignment employment. Estimation is by OLS.

6. Replication of Mincome Labor Supply Results for Public Use File

Table A.6.1
Estimated Treatment Effects: Labor Supply, "Minc4" Public Use File

	Women		Men	
	Hours Worked	Employed	Hours Worked	Employed
Treatment Group	-48.35*** (19.67)	-.065* (.039)	-27.47 (19.00)	-.036 (.024)
Mean Dep. Variable (Controls)	212.2	.449	505.1	.846
Coefficient in annual hours	145	-	90	-
Number of Individuals	418		420	
Sample Size	3159		3162	

NOTES—Standard errors are in parentheses, and are clustered on the individual. Statistical significance denoted by *** for the 1% level, ** for the 5% level and * for the 10% level. All regressions include fixed effects for the experimental stratification (income cell for Mincome), dummies for survey/time, in addition to demographic controls and pre random assignment employment. Estimation is by OLS.