Evaluation Plan: Minneapolis Guaranteed Basic Income Pilot

Andrew Goodman-Bacon, Vanessa Palmer, and Ryan Nunn
Federal Reserve Bank of Minneapolis

Abstract: This document outlines the pre-analysis plan for the Federal Reserve Bank of Minneapolis' analysis of the City of Minneapolis' Guaranteed Basic Income (GBI) Pilot. The study contains 530 households drawn from 9 ZIP codes in Minneapolis. Treated households receive $500 per month for 24 months. Control households receive $150 for completing surveys at 6 months, 12 months, and 24 months. Participants were randomly assigned to treatment within 8 strata defined by poverty, presence of children, and two ZIP code groups. Early enrollment data showed significant differential attrition. The plan addresses biases from selective attrition by (a) conditioning on age, baseline income, and education (b) controlling for the baseline outcome variable and (c) estimating treatment/control contrasts for the change in outcomes (difference-in-differences).

Evaluation Plan:
Minneapolis Guaranteed Basic Income Pilot[*]

[*] *We gratefully acknowledge support from Alex Albright, Michael Anderson, Emily Beam, Tyler Boesch, Jake Bowers, JP Bruno, Dalia Ghanem, Natalie Gubbay, Jason Kerwin, Pedro Sant'Anna, Abbie Wozniak, and four anonymous reviewers.*

I. Background

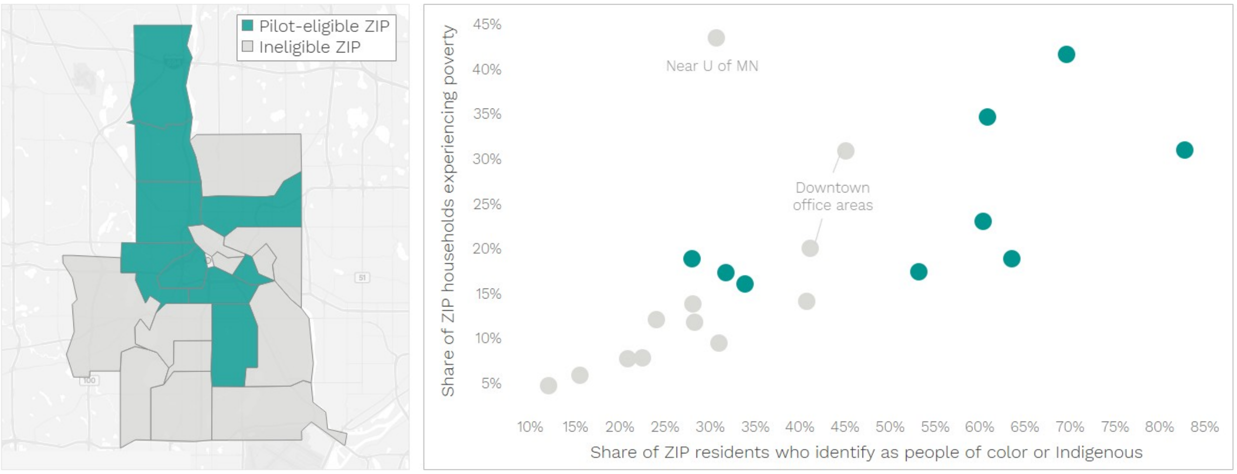
1. Motivation and Funding

In July of 2021, the City of Minneapolis (the City) announced plans to pilot a guaranteed basic income (GBI) program. Resourced by $3M in federal American Rescue Plan Act (ARPA) funds through the Coronavirus State and Local Fiscal Recovery Funds (SLFRF) program, the pilot aims to distribute unconditional financial support to economically vulnerable City residents affected by the Covid-19 pandemic, and to assess the impacts of this support on a targeted range of socioeconomic indicators. Total funding for the GBI initiative represented approximately 1% of the City’s $271M SLFRF aid package.

The pilot channels Covid-19 relief funds to low- and moderate-income Minneapolis residents of nine ZIP codes. These ZIP codes were selected on the basis of evidence of systemic barriers to economic opportunity, including ZIP-level prevalence of poverty (Figure P1). Additional participant eligibility criteria are detailed in Section II.

The Federal Reserve Bank of Minneapolis (FRB-MPLS) serves as the independent program evaluator for the City's GBI pilot. Through this research relationship, FRB-MPLS advances its study of policies affecting labor market dynamics in low- and moderate-income communities.

Figure P1. Characteristics of GBI Pilot-Eligible ZIP Codes



Source: Federal Reserve Bank of Minneapolis calculations using City of Minneapolis shapefile, U.S. Census Bureau TIGER/Line shapefiles, and U.S. Census Bureau American Community Survey five-year file, 2015–2019.

2. Project Roles and Team

The City of Minneapolis leads the GBI pilot's design, oversight, and implementation. The Federal Reserve Bank of Minneapolis leads evaluation.

Mark Brinda, Ph.D.
Workforce Manager, Employment and Training
Community Planning and Economic Development—Economic Policy and Development
City of Minneapolis

Andrew Goodman-Bacon, Ph.D.
Senior Research Economist
Research Division—Opportunity and Inclusive Growth Institute
Federal Reserve Bank of Minneapolis

Katherine Lim, Ph.D. *(former)*
Economist
Community Development and Engagement—Applied Research
Federal Reserve Bank of Minneapolis

Jeremy Lundborg
Special Projects Manager
Community Planning and Economic Development—Economic Policy and Development
City of Minneapolis

Andrea Naef
Assistant City Attorney
Office of the City Attorney
City of Minneapolis

Ryan Nunn, Ph.D.
Assistance Vice President
Community Development and Engagement—Applied Research
Federal Reserve Bank of Minneapolis

Vanessa Palmer, Sc.M.
Data Scientist
Community Development and Engagement—Applied Research
Federal Reserve Bank of Minneapolis

Data privacy is a priority. By design, any contact with pilot participants is by the City of Minneapolis and its community-based designees. All information used for evaluation purposes is stripped of personally identifiable elements before it is securely transferred by the City to the evaluation team at the Federal Reserve. Data movement, use, and storage are governed by a legally binding agreement between the two entities.

3. Project Timeline

Italics indicate a stage that is in the future as of this writing. An asterisk (*) indicates a tentative stage.

July 2021
The City of Minneapolis (“the City”) publicly announces its GBI pilot.

Fall 2021
The City plans and communicates about the pilot, while collaborating with the Federal Reserve Bank of Minneapolis (“the evaluation team”) to design the evaluation.

December 2021
The GBI pilot interest form opens. Individuals interested in participating in the pilot share contact details and preliminary eligibility-related information.

January 2022
The City consolidates duplicate interest form submissions and screens for initial eligibility. The evaluation team performs data validation followed by a randomization process to select a subset of preliminarily eligible individuals that the City will invite to complete the baseline survey.

March 2022
Invited, preliminarily eligible individuals complete the baseline survey.

April 2022
The evaluation team performs data validation followed by a randomization process to sort baseline survey respondents into groups to be invited to join as payment (treatment) or survey (control) participants.

April–June 2022
The City and its designees at two community-based organizations (CBOs) conduct intake processes for payment participants until the payment group reaches capacity.

June 2022
Monthly cash transfers to confirmed payment participants begin. The City and its CBO designees begin intake for survey group participants.

December 2022–January 2023
At six months since payments began, all participants (treatment and control) are asked to complete a follow-up survey.

May 2023
The evaluation team produces an interim update communicating key findings to date.

June–July 2023
At 12 months since payments began, all participants (treatment and control) are asked to complete a follow-up survey.

**December 2023–January 2024*
At 18 months since payments began, all participants (treatment and control) are asked to complete a follow-up survey.

March 2024
The evaluation team produces an interim update communicating key findings to date.

June 2024
Monthly cash transfers to payment participants stop.

June–July 2024
At 24 months since payments began, all participants (treatment and control) are asked to complete a follow-up survey.

Fall 2024
The evaluation team produces a final report communicating key findings.

**June–July 2025*
At 36 months since payments began (12 months since payments stopped), all participants (treatment and control) are asked to complete a follow-up survey.

II. Structure and Planning of the Evaluation

1. Treatment, Control, and Eligibility Parameters

The treatment of interest is randomized assignment to a $500 unconditional monthly cash transfer for a period of 24 months.

The outcomes of interest are a targeted range of socioeconomic indicators detailed in Section III.

To estimate the associations between the treatment and the outcomes of interest, the evaluation makes comparisons between the treatment group ("payment participants") and a control group not receiving the treatment ("survey participants"). To maximize the comparability of the two groups, the same eligibility criteria and verification steps were applied to both.

Aside from its necessity for evaluation purposes, eligibility verification also carried legal importance, as the pilot applies federal American Rescue Plan Act (ARPA) Coronavirus State and Local Fiscal Recovery Funds (SLFRF) program funds with the express intent of addressing the disproportionate pandemic-related economic harms experienced by low- and moderate-income households.[†]

Thus, at the time of their eligibility verification, each participant:

- Was a Minneapolis resident of one of nine eligible ZIP codes: 55403, 55404, 55405, 55407, 55411, 55412, 55413, 55430, or 55454
- Was 18 years of age or older as of January 1, 2022
- Attested to having a status which would make them eligible to receive federally-funded benefits[‡]
- Attested to having experienced economic challenges related to the pandemic
- Provided documentation or attestation of a 2021 annual household income less than or equal to 50% of that year's area median income (Table P1)

Once verified as eligible, participants can lose eligibility only by moving out of the City of Minneapolis.

Table P1. GBI Pilot Household Income Eligibility Thresholds by Household Size

1	$36,725	**5**	$56,646	**9**	$73,425
2	$41,975	**6**	$60,842	**10**	$77,625
3	$47,225	**7**	$65,038	**11**	$81,825
4	$52,450	**8**	$69,234	**12**	$86,025

Source: U.S. Department of Housing and Urban Development (2021), https://www.huduser.gov/portal/datasets/il.html

2. Recruitment

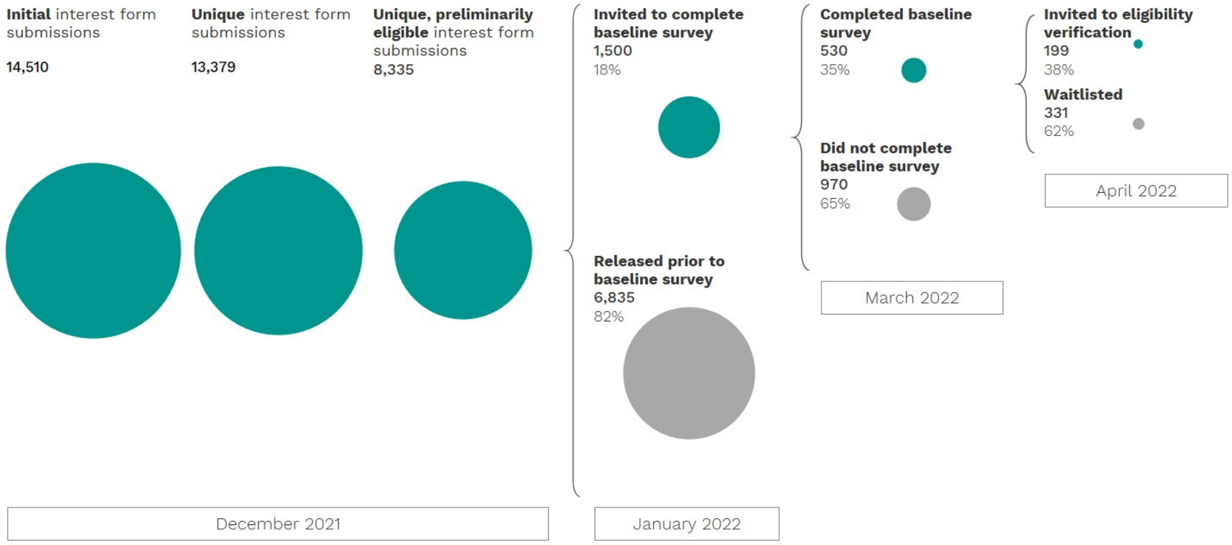
[†] https://home.treasury.gov/policy-issues/coronavirus/assistance-for-state-local-and-tribal-governments/state-and-local-fiscal-recovery-funds

[‡] Eligible statuses included but were not limited to United States citizens or nationals, permanent residents, and refugees. Ineligible statuses included international students.

In fall 2021, the City of Minneapolis conducted outreach to communicate the objectives of the pilot and recruit prospective participants from the community. Outreach activities included partnerships with community-based organizations; print and online advertisements; interviews in local media; and a landing page on the City website.

An initial interest form was open for the month of December 2021. With the aim of minimizing barriers to entry into the applicant pool, information collected on the interest form was intentionally limited. The first panel (three circles) in Figure P2 document the initial phase of recruitment. The City received 14,510 submissions, of which the City determined 13,379 were unique. Of these, the City removed applicants who did not qualify based on geographic, age, and residency eligibility criteria. 8,335 unique, preliminarily eligible interest form submissions remained. The City removed all personally identifiable information and assigned a unique, anonymized identifier before then passing the records to the evaluation team.

Figure P2. Stages of Recruitment



With a limited, 200-participant capacity in the payment group, care was taken during the remaining selection stages to limit unnecessary requests of community members' time and to minimize the potential for harm through a protracted, high-stakes process.

The evaluation team randomly sequenced the 8,335 anonymized, preliminarily eligible interest form submissions and the top 600 were invited by the City via email to complete the baseline survey. After two weeks, the next 900 were invited to complete the baseline survey, for a total of 1,500 invited applicants, shown in the second panel circles of Figure P2.

The baseline survey was administered online from 1 March to 29 March 2022. It included a Data Practices Advisory that prospective respondents were required to affirm having read and understood before submitting information. Eligibility-related items were at the beginning of the survey; responses indicating ineligibility caused text to appear advising respondents accordingly. The written version of the baseline survey was in English, with prominent advisory text directing respondents to live City translation support in six additional languages: Spanish, Hmong, Somali, Lao, Oromo, and Vietnamese. The third panel of Figure P2 shows that of the 1,500 individuals invited to complete the baseline survey, 530 completed it, including three who reported having used language/translation support.

To these 530 respondents, the evaluation team applied a final randomization process detailed in the next section. An intake stage followed.

3. Baseline Survey Instrument

Items for the baseline survey instrument were primarily selected from existing, publicly available sources. This was done to facilitate external comparisons, both to appropriately contextualize Minneapolis results and to contribute to the growing body of evidence on basic income programs.

A community focus group in fall 2021 facilitated by an external community-based organization provided input on survey design considerations. A second community focus group provided field testing and feedback in January 2022.

In pre-testing, typical baseline survey completion time was between 20 and 30 minutes.

The baseline survey included 169 possible items, including demographic items, operational items, and conditional items for capturing details regarding preceding responses of “other.” Of these, 60 items were reserved for capturing information for up to ten household members in addition to the respondent.

The baseline survey instrument will be made publicly available on the Federal Reserve Bank of Minneapolis website.

4. Randomization and Intake

a. Sequencing

Following the completion of the baseline survey, an intake process was conducted individually for each participant. Intake consisted of eligibility verification and, for individuals invited to the treatment group:

(1) The suggestion (with supporting written resources) to consult benefits counselor(s) as needed regarding the potential impact of GBI payments—considered taxable income—on other forms of income-based assistance (e.g., Supplemental Nutrition Assistance Program support)[§] before making a final decision about participation, and

(2) Enrollment into direct deposit (City-preferred) or the collection of details for an alternative payment method.

Thus, intake represented a significant investment of time by City staff and designees in email and phone contacts, meetings, and document reviews. Because

[§] The Minnesota Department of Human Services announced in November 2022 that GBI payments were to be excluded from eligibility considerations in the programs it oversees. This exemption was not yet in place at the time of intake for this pilot, though treatment participants who later reported impacts were advised.

the pilot's funding source was time-delimited**, this necessitated a phased approach such that treatment participants' intake could be completed first and monthly payments could begin.

From the sample of 530 individuals who responded to the baseline survey, the evaluation team used stratified random sampling to assign each respondent to one of two lists: an immediate intake list to become a treatment participant, or a sequenced waitlist from which participants were drawn for intake as needed until treatment group capacity was reached. It is important to note that the stratification strategy did *not* give priority to any particular demographic group (e.g., families with children). Rather, it removed the possibility that (by random chance) members of a particular demographic group were disproportionately selected for payment. In so doing, it also facilitated subsequent evaluation of treatment effects.

b. Strata

We defined strata based on the 8 ($¿2^3$) combinations of the following dummy variables using individuals' baseline survey responses:

$$kids_i=\begin{cases}1\,if\,household\,contains\,anyone\,under\,age\,18\\ 0\,otherwise\end{cases}$$

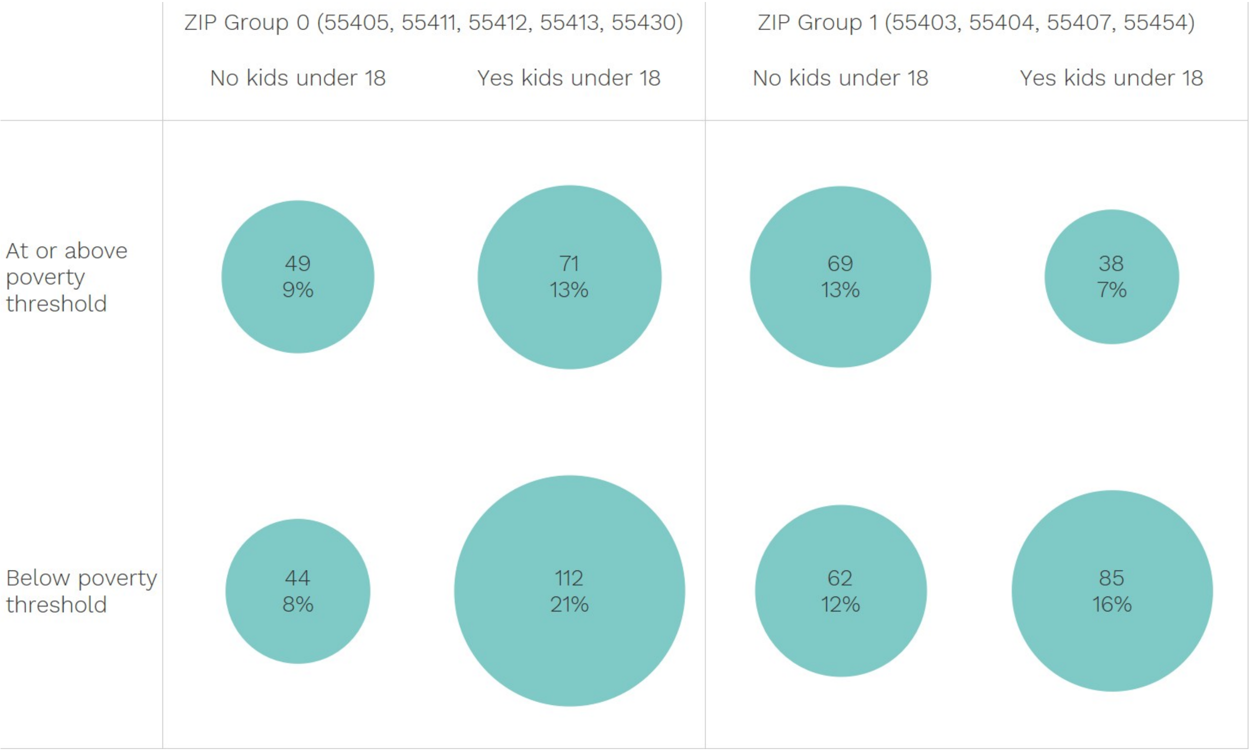
$$zip\,1_i=\begin{cases}1\,if\,household\,lives\in zip\,codes\,55403,55404,55407\vee 55454\\ 0\,if\,household\,lives\in zip\,codes\,55405,55411,55412,55413\vee 55430\end{cases}$$

$$pov_i=\begin{cases}1\,if\,self-reported\,household\,income\,is\,below\,the\,2021\,Federal\,Poverty\,Level\\ 0\,otherwise\end{cases}$$

The ZIP code groups roughly split respondents into neighborhoods south of downtown or in central Minneapolis ($zip\,1_i=1$) versus in North or Northeast Minneapolis ($zip\,1_i=0$). Between 40 and 90 households fell in each stratum, as shown in Figure P3.

** Coronavirus State and Local Fiscal Recovery Funds (SLFRF) program funds must be obligated by December 31, 2024. https://www.govinfo.gov/content/pkg/FR-2022-01-27/pdf/2022-00292.pdf

Figure P3. Distribution of Sampled Households Across Strata



c. Assignment Probabilities and Mechanism

The City's goal was to enroll a total of 200 individuals in the treatment group. From a practical standpoint, the evaluation team needed to provide a clear post-randomization list to City staff who would then offer treatment to individuals and confirm their eligibility and desire to participate. Individuals initially offered treatment and found to be ineligible or who opted out were to be replaced, but it was not feasible to make this step adaptive to realized attrition by strata, i.e., to adjust the waitlist in response to attrition.

We therefore used the following procedure to choose individuals for invitation to treatment, determine eligibility and participation, and, if necessary, choose additional individuals until a total of 200 were enrolled in the treatment group:

Step 1. Assign a random number to all sample members and rank them within their stratum. Denote household i's ranking in stratum j by r_i^j.

Step 2. Define the initial number of treatment units in stratum j as the top $N_j^T = round\left(200 \times \frac{N_j}{530}\right)$ individuals in that stratum. Therefore, initial treatment allocation is $D_i = 1\{r_i^j \leq N_j^T\}$. Given the sample sizes and strata distributions, this yielded a list of 199 individuals initially randomized to be offered treatment.

Step 3. Define a random number for the remaining 331 individuals and rank them by it. Denote individual i's ranking in this waitlist r_i^C. Assign the first individual in the waitlist to the 200th treatment spot.

Step 4. Contact individuals randomized to treatment invitation to begin intake. If an individual is ineligible or does not wish to participate, replace them with the next individual in the waitlist, i.e., the individual with the lowest remaining r_i^C draw.

In practice, randomization followed steps 1-3, but step 4 differed slightly. The City and its community-based organization (CBO) designees began intake for the treatment group in April 2022. By June 2022, City and CBO staff determined that some individuals randomized to treatment invitation were ineligible, did not wish to participate, or had not yet responded after multiple successive emails, phone calls, and text messages. To start the intake process for their replacements, City and CBO staff began inviting individuals from the waitlist into the treatment group. In principle this could have generated more than 200 treatment participants, but in practice it did not.

In total, 200 individuals enrolled in the treatment group and as of March 2023, 134 individuals had completed intake as control participants. Intake will continue on a rolling basis for controls, using the same eligibility criteria. Of note: The interim results in this report include information only from participants who have completed the intake process as of the time of writing (N = 334).

5. Follow-up Survey Waves

a. Instruments

The six-month follow-up survey is identical to the baseline survey except:

- We did not re-collect demographic or annual income information.
- In collaboration with the City legal team, we updated the Data Practices Advisory to advise respondents that current address information would be used to confirm continuing eligibility.
- We added an open-ended item about changes in financial circumstances over the past six months.
- For treatment group participants, we added items about their experiences with the pilot, including open-ended items.
- We adjusted a pair of questions about income sources to use a yes/no indicator and categories rather than continuous amounts.

We anticipate minimal or no changes to subsequent survey instruments.

Survey instruments will be made publicly available on the Federal Reserve Bank of Minneapolis website.

b. Administration

Six-month surveys were administered online from 27 December 2022 to 27 January 2023. During this window, City staff and designees followed up with potential respondents via email and, later, by phone call.

We anticipate a similar data collection cycle in subsequent survey rounds.

All control participants who have completed the intake process receive $150 for completing each survey wave.

c. Data Linkages

In an effort to prioritize a successful implementation of core pilot elements, we did not seek consent from participants to access administrative data held by other agencies or programs.

III. Pre-Analysis Plan: Minneapolis Guaranteed Basic Income Pilot

To understand if GBI affects its recipients, in what ways, and by how much we will compare average outcomes for households that received payments, the **treatment group**, to households that did not receive payments, the **control group**. Making this comparison is the simplest part of our analysis. It only involves calculating averages from the data we collect. The biggest challenge in this study is to discern whether we can interpret treatment/control comparisons as a **causal effect** of GBI. This requires assumptions about whether the experience of control households reflects what would have happened to treatment households had they not received GBI.

The plan outlined in this section therefore links three pieces of our analysis:

- the kind of causal effect we would like to estimate (parameter of interest)
- the statistical assumption we must believe in order for specific comparisons in the data to be informative about that parameter (identifying assumption)
- how we make those comparisons (estimator)

We describe the procedures and methods used to measure outcome variables; define the treatment effect, parameters of interest, and statistical assumptions required to identify them; characterize attrition and non-response; estimate treatment effects; perform statistical inference; evaluate sensitivity of the findings; and report results.

1. Outcomes

Our detailed survey instrument gathers information on outcomes that might change in response to GBI, but its wide scope presents three challenges:

- **Interpretability**. We do not expect readers or stakeholders to draw clear conclusions from hundreds of estimated effects.
- **Imprecision**. If GBI does affect an outcome only for a handful of families, then our study may not be able to detect it. Detailed survey questions can create “rare outcomes” that are statistically difficult to evaluate.
- **False discoveries**. For every outcome we study, we face a random chance of incorrectly concluding that GBI has some effect. The more statistical tests we perform, the higher the chance that we mistakenly report a false discovery.

Consider an important concept like housing stability that lacks a single validated measure in wide use. If guaranteed income makes it easier for families to afford rent, then they may worry less about having to move, they may *actually* move less often, and they may not have to double up in crowded housing with family or friends.

Suppose that GBI does causally affect each of these outcomes, but only a few households respond to each one. In that case, the treatment group’s average outcomes may not change very much in response to GBI and our study may not be able to distinguish these changes from zero (imprecision). Alternatively, suppose that GBI has no effect on these outcomes. By testing for an effect of GBI on each one, we increase the chance that at least one of our findings is an error (false discovery). Lastly, each of these measures contributes to the concept of housing stability. Estimating separate GBI effects for each variable makes it difficult to learn whether GBI affects this broader type of outcome (limiting interpretability).

To address all three challenges, we combine related measures into outcome indices. We then estimate a single treatment effect on each index. This is easier to interpret than a collection of separate estimates and more closely reflects the broad concepts that the study seeks to evaluate. It makes our estimates more precise since it accounts for the fact that families may respond to different *facets* of a given concept. And it protects against false discoveries because we only have one chance for a random error, rather than many.

We combine our survey variables into 10 indices, each representing a different outcome domain. We code all component variables so that positive values indicate an “improvement” or “more” of a given concept. We first describe the components of each index and then our method for aggregation.

a. Index Definitions

For further details about specific survey items, please see the survey instruments in the Appendix.

i. Labor Supply (INDEX_LSUPP)

This index uses six concepts—constructed from 12 variables—to measure how much participants work:

- Respondent worked in the last month (WORK_ANYMONTH==“Yes”)
- Respondent employed in the week before the survey (WORK_ANYWEEK==“Yes” | WORK_TEMPABS==“Yes” | WORK_UNPAYFAM==“Yes”)
- Respondent in the labor force in the week before the survey ((WORK_ANYWEEK==“Yes” | WORK_TEMPABS==“Yes” | WORK_TEMPLAY==“Yes” | WORK_UNPAYFAM==“Yes”) | (WORK_SEARCH==“Yes” & WORK_AVAIL==“Yes”))
- Respondent working full-time (WORK_MAINFTPT==“Usually full-time” | WORK_MAINHRWK>=35)
- Respondent had multiple jobs in the week before the survey (WORK_ANYADDLWK==“Yes”)
- Respondent total usual weekly hours worked at all jobs (WORK_MAINHRWK*(WORK_MAINHRWK~=“DKR”) + WORK_ADDLHRWK*(WORK_ADDLHRWK~=“DKR”))

ii. Housing Stability (INDEX_HSTABLE)

This index uses nine concepts to measure how stable a participant’s housing situation is:

- Respondent lives in a house or apartment (HOUSE_CURR==“House” | HOUSE_CURR==“Apartment”)
- Respondent/household owns or rents their housing (HOUSE_OWNRENT==“Owned by you or someone in your household with a mortgage or loan” | HOUSE_OWNRENT==“Owned by you or someone in your household free and clear” | HOUSE_OWNRENT==“Rented by you”)
- Household did not experience difficulty affording housing payment in the previous six months (HOUSE_AFFORD6==“Never”)
- Household was not late on rent or mortgage in the previous six months (HOUSE_LATE6==“No”)
- Respondent does not feel that housing is overcrowded (HOUSE_CROWD==“No”)
- Persons per bedroom in respondent’s housing unit is below overcrowding measure suggested by U.S. Department of Housing and Urban Development (constructed from HOUSE_BEDROOMS, HH_COUNT, and HH_OTHCOUNT; total persons in housing unit / bedrooms in housing unit ≤ 2)
- No household experiences of housing instability over the previous six months (HOUSE_INST6==“None of the above”)
- No household worry about a forced move over the previous six months (HOUSE_MOVEFWORRY6==“Never”)
- Respondent/household did not experience a forced move in the previous six months (HOUSE_MOVE6==“Yes" & HOUSE_MOVE6REAS==“Forced”)

iii. Financial Security (INDEX_FINSEC)

This index uses 11 concepts to measure a household’s financial stability:

- Self-reported overall financial situation (FIN_OVERALL==“Living comfortably” or FIN_OVERALL==“Doing okay”)
- Not getting income from sources other than working (i.e., public assistance, family or friends, or other sources) (FIN_HOUSEINCNOWORKYN==“No”)
- No charity food assistance (FIN_SUPPFOOD6==“No”)
- No charity financial support (FIN_SUPPFINORG6==“No”)
- No family financial support (FIN_SUPPFINPERS6==“No”)
- Provide financial support for others (FIN_SUPPFINPERSPROV==“Yes”)
- Any precautionary saving (FIN_SAVE3MO==“Yes”)

- Could cover three months' expenses (FIN_COVER3MO==“Yes”)
- Could cover a $400 emergency expense (FIN_400HOW~=“I wouldn't”)
- Able to pay all bills (FIN_PAYBILL==“Able”)
- Behind on debt (FIN_DEBTBEHIND==“No”)

iv. Well-Being (INDEX_WB)

This index uses three concepts to measure respondents' self-reported general well-being. Item scores are reverse-coded from the scales outlined in the sources from which they originated:

- General health (WELL_HEALTH)
- Overall happiness (WELL_HAPPY)
- General life satisfaction (WELL_SATISFIED)

v. Food Security (INDEX_FOODSEC_SECURE)

This binary outcome is constructed from the US Department of Agriculture (USDA)'s six-item short form of its food security survey.[††] Respondents are asked a series of questions referring to the last 30 days. Results are aggregated according to USDA's scoring guidance. Scores of zero or one (on a scale of six) represent a screening result of “high or marginal food security” in a household, and are considered here food secure.

- Thought food wouldn't last (FOOD_LAST)
- Couldn't afford balanced meals (FOOD_BAL)
- Skipped meals or cut the size of meals (FOOD_SIZESKIP)
- Number of days skipped/cut size (FOOD_SIZESKIPFREQ)
- Ate less than you should (FOOD_EATLESS)
- Did not eat despite feeling hungry (FOOD_NOTEAT)

vi. Psychological Wellness (Kessler 10) (INDEX_K10)

This outcome is based on one version of a common, clinically predictive screening tool for psychological distress.[‡‡] Items ask how often respondents experienced various feelings in the last 30 days, with responses on a five-element scale ranging from “none of the time” (scored as one) to “all of the time” (scored as five). Responses are then summed. Consistent with screening guidance and other evaluations, we use a score of 20 as a threshold under which a respondent is considered not to have mental health problems.

- Tired out for no good reason (HEALTH_KTIRED)

[††] https://www.ers.usda.gov/topics/food-nutrition-assistance/food-security-in-the-u-s/survey-tools/

[‡‡] https://pubmed.ncbi.nlm.nih.gov/12214795/

- Nervous (HEALTH_KNERV)
- So nervous that nothing could calm you down (HEALTH_KCALM)
- Hopeless (HEALTH_KHOPE)
- Restless or fidgety (HEALTH_KREST)
- So restless you could not sit still (HEALTH_KSTILL)
- Depressed (HEALTH_KDEPR)
- So depressed that nothing could cheer you up (HEALTH_KCHEER)
- Everything was an effort (HEALTH_KEFFORT)
- Worthless (HEALTH_KWORTH)

vii. Housing Quantity (INDEX_HQUANT)

This index uses four concepts to measure the amount of housing that participants consume:

- Respondent/household owns their housing (HOUSE_OWNRENT==“Owned by you or someone in your household with a mortgage or loan” | HOUSE_OWNRENT==“Owned by you or someone in your household free and clear”)
- Rent or mortgage payment (HOUSE_COSTMO)
- Number of bedrooms (HOUSE_BEDROOMS)
- Respondent made a planned move in the previous six months (HOUSE_MOVE6==“Yes” & (HOUSE_MOVE6REAS==“Wanted…” | HOUSE_MOVE6REAS==“Household changed…”))

viii. Use of Low-Cost Credit (INDEX_CRED)

This index uses five concepts to measure participants’ households’ avoidance of higher-cost sources of credit:

- No non-bank money order (FIN_CREDMONORD~=“Yes”)
- No non-bank check cash (FIN_CREDCASH~=“Yes”)
- No payday loan (FIN_CREDPAYDAY~=“Yes”)
- No pawn shop/auto title loan (FIN_CREDLOAN==“Yes”)
- No respondent unpaid credit card balance (FIN_CC==“No” | FIN_CCBALFREQ6==“Never”)

ix. Healthcare Utilization (INDEX_HUSE)

This index uses six items to measure the extent to which respondents reported a household member using various healthcare services in the previous six months:

- Prescription medicine (HEALTH_UPRESC6)
- Visit to a doctor or specialist (HEALTH_UDOC6)
- Mental health care or counseling (HEALTH_UMENT6)
- Dental care (HEALTH_UDENT6)
- Follow-up care (HEALTH_UFOLL6)
- Emergency room (HEALTH_UEROOM6)

x. Healthcare Access (INDEX_HACC)

This index uses six items to measure the extent to which respondents reported—in the previous six months—a household member having needed various healthcare services, but having gone without due to financial constraints. A response of "no" is coded positively.

- Prescription medicine (HEALTH_NPRESC6)
- Visit to a doctor or specialist (HEALTH_NDOC6)
- Mental health care or counseling (HEALTH_NMENT6)
- Dental care (HEALTH_NDENT6)
- Follow-up care (HEALTH_NFOLL6)
- Emergency room (HEALTH_NEROOM6)

b. Aggregation Method

To calculate these indices we follow Anderson's (2008) inverse covariance weighting approach. (The food security index and Kessler 10 scale use externally comparable aggregation methods—detailed above—instead of the one described here.) Variables that are strongly correlated with other components do not add new information and are given little weight, while variables with low correlations to other measures represent more new information are weighted more heavily. Intuitively, a key advantage of this approach is that it makes it less important whether several closely related questions were asked, or only one question on a given sub-topic. If equal weights are applied to all components, then the decision to ask multiple related questions is very influential for the interpretation of results. By contrast, the Anderson (2008) approach more reliably approximates the underlying concept we are attempting to measure in each instance.

Consider domain k with component variables y_{ilt} with $l=1,\ldots,K$. $t=0$ for baseline measures. The following four steps create the inverse-covariance-weighted index:

Step 1. Standardize each variable by its mean and standard deviation in the full baseline sample:

$$z_{ilt}=\frac{y_{ilt}-\overline{y}_{l0}}{\hat{\sigma}_l^0}$$

Where $1\{y_{il0} \neq .\}$ Is an indicator for household i responding to item k at baseline, $\overline{y}_{l0} \equiv \frac{\sum_{i=1}^{N} y_{il0} 1\{y_{il0} \neq .\}}{\sum_{i=1}^{N} 1\{y_{il0} \neq .\}}$, and $\hat{\sigma}_l^0 \equiv \sqrt{\frac{\sum_{i=1}^{N} (y_{il0} - \overline{y}_{l0})^2 1\{y_{il0} \neq .\}}{\left(\sum_{i=1}^{N} 1\{y_{il0} \neq .\}\right) - 1}}$. Outcomes from all follow-up waves will be standardized using baseline means and standard deviations.

Step 2. Calculate the inverse covariance matrix of all the z_{ikt} variables:

$$\Sigma_K^{-1} \equiv \begin{bmatrix} 1 & \dots & \hat{\sigma}_{1K} \\ \vdots & \ddots & \vdots \\ \hat{\sigma}_{K1} & \dots & 1 \end{bmatrix}^{-1} = \begin{bmatrix} c_{11} & \dots & c_{1K} \\ \vdots & \ddots & \vdots \\ c_{K1} & \dots & c_{KK} \end{bmatrix}$$

Where $\hat{\sigma}_{lm} \equiv \frac{\sum_{i=1}^{N} z_{ilt} 1\{z_{ilt} \neq .\} z_{imt} 1\{z_{imt} \neq .\}}{\left(\sum_{i=1}^{N} 1\{z_{ilt} \neq .\} 1\{z_{imt} \neq .\}\right) - 1}$. Each covariance in Σ_K is based on as many observations as are non-missing for both variables.[§§]

Step 3. For component variable l define $\tilde{w}_k$ as the sum of the elements in the lth row of Σ_K^{-1}:

$$\tilde{w}_l = \sum_{m=1}^{K} c_{lm}$$

Step 4. For unit i create an index that equals an average of that unit's non-missing variables in domain k weighted by $\tilde{w}_l$:

$$Y_{ikt} \equiv \sum_{l=1}^{K} \frac{\tilde{w}_l 1\{z_{ilt} \neq .\}}{\sum_{m=1}^{K} \tilde{w}_m 1\{z_{imt} \neq .\}} z_{ilt} = \sum_{l=1}^{K} w_{li} z_{ilt}$$

For more detail see appendix A in Anderson (2008). We do not use this method for variables that come from existing approaches to aggregation, such as the Kessler 6 and food security measures.

c. Other Outcomes

Some outcomes are of independent interest outside of their broad domain. For example, current employment is a commonly discussed outcome in many GBI

[§§] Many statistical packages calculate covariance matrices using the set of observations that are available for *all* components of the matrix Σ_K, but we explicitly use all available information to calculate each covariance. Since we have baseline data for all participants observations would only differ across entries in this matrix because of item non-response on the baseline survey. To check whether this approach influence our outcome measures we will also calculate the indices with weights based on covariance matrices that use the same set of observations for all elements and report a correlations between the two measures for each domain.

debates. Below when we discuss statistical methods to protect against false discoveries, we draw a distinction between more stringent methods we use to analyze selected index variables (i-vi) defined above and less stringent methods that we use in an “exploratory analysis” of the other indices (vii-x) and certain component variables. Outcomes in the exploratory analysis will include:

i. INDEX_HQUANT, INDEX_CRED, INDEX_HUSE, and INDEX_HACC

ii. Respondent employed in the week before the survey (WORK_ANYWEEK==“Yes” | WORK_TEMPABS==“Yes” | WORK _UNPAYFAM==“Yes”)

iii. Respondent works multiple jobs (WORK_ANYADDLWK==“Yes”)

iv. Respondent’s household could cover a $400 expense (FIN_400HOW~=“I wouldn’t”)

v. Respondent reported or implied hourly wage (among workers) (constructed from WORK_MAINHRWK, WORK_MAINEARNPERIOD, WORK_MAINEARNAMT, WORK_ADDLHRWK, and WORK_ADDLEARNWK)

vi. Respondent’s household provides financial support to others (FIN_SUPPFINPERSPROV==“Yes”)

vii. Respondent’s household not flagged on common housing instability screener (HOUSE_INST6==“None of the above”)

viii. Respondent transportation access (WELL_TRANS==“Always” | WELL_TRANS==“Often”)

ix. Respondent school/training attendance (EDUC_SCHATT6==Yes | (EDUC_TRAINATT6==“Yes” & (EDUC_TRAINCOMP6==“Yes” | EDUC_TRAINCOMP6==“No, but still attending”))

x. Points in the cumulative distribution of annual household income (asked only at annual intervals) (constructed from FIN_HOUSEINC)

2. Target Parameters and Identifying Assumptions

We begin our analysis plan by defining the kinds of causal effects we want to estimate, the population to which these effects apply, and the assumptions under which we can estimate them. This provides clarity about what can and cannot be learned from this study.

Next we use our full baseline survey results to describe study participants in more detail, including their outcomes before receiving GBI payments. Because we based our survey instrument on existing public datasets, we will be able to compare study participants to broader samples as well as to samples from other GBI studies.

We also evaluate the likelihood that our study, as implemented, can generate plausible causal estimates. This is called internal validity. We first test whether our randomization produced comparable treatment and control groups prior to GBI payments. We also use information on response patterns to the 6-month survey to quantify and describe the sample of *respondents:* those who provided post-randomization data that we can use in our analysis.[***] Selective attrition means that different kinds of participants provided follow-up data in the treatment and control groups. This can mean that GBI treatment is no longer random among respondents, despite successful randomization of the full population to treatment and control groups.

a. Study Population

The **population** to which our study results apply is defined by three conditions:

a) Applied to the GBI pilot,

b) provided valid information and appeared initially eligible (see section II.1),

c) would have filled out the baseline survey upon request (see section II.4).

The application process and initial vetting of applications (based on duplications, address, age, and self-reported income) define (a) and (b). Our random selection of 1,500 applications to receive baseline surveys preserves the population characteristics of (a) and (b). Responses to the baseline survey (530 out of 1,500) generate a sample that satisfies (c) (see Figure P2).

b. Notation

We use the following notation to define our sample, outcomes, causal effects of interest, and identifying assumptions:[†††]

- S_i denotes **s**ampled households from the population defined by (a)-(c).
- R_{it} denotes **r**esponse status at follow up. $R_{it}=1$ if respondent i formally verified their eligibility and provided enough information to receive payments (either GBI or survey payments) and responded to the survey in period t.
- D_i denotes randomly assigned treatment status for unit i.
- $G_i=1,\dots,8$ denotes household i's stratum and $g_{j(i)}=1\{G_i=j\}$ is a dummy variable equal to one if household i is in stratum j.
- $Y_{it}(1)$ and $Y_{it}(0)$ denote treated and untreated potential outcomes for unit i at time t. $t=0$ denotes the baseline period.
- $R_{it}(1)$ and $R_{it}(0)$ denote treated and untreated potential response status for unit i at time t. $t=0$ denotes the baseline period.

c. Target Parameters

[***] Dropping out of a study is called *attrition* or *non-response.* When respondents answer some survey questions but not others it is called *item non-response*. Since attrition will change over time, we will re-evaluate it at each follow up wave. As we worked with the city to administer the 6-month survey, non-response patterns became clear. Therefore, we include follow-up information on attrition (but not outcomes) in our baseline evaluation of internal validity and when choosing our estimation approach.

[†††] When we define our parameters of interest and identifying assumptions we use the expectations operator, $E[\cdot]$. For a theoretical population, the expectation is the analog of the sample average. We also often write conditional expectations such as $E[a \vee x=b]$, which is the average of the variable a among the sub-population for whom another variable, x, equals b.

Potential outcomes define the concept of a treatment effect for every unit.[‡‡‡] Unit i's treatment effect equals the difference between its outcomes in the state of being treated and the state of being untreated: $Y_{it}(1)-Y_{it}(0)$.[§§§]

We are interested in **average treatment effect parameters**.[****] Depending on the nature of attrition, our design can potentially estimate two types of average causal effects:

- **Average treatment effect (*ATE*)** of monthly GBI payments relative to survey completion payments among the population satisfying (a)-(c):

$$\tau_{ATE,t} \equiv E\left[Y_{it}(1)-Y_{it}(0)\right]$$

$$\text{¿}\sum_{j=1}^{8} n_j E\left[Y_{it}(1)-Y_{it}(0) \vee G_i = j\right]$$

$$\text{¿}\, E\left[\tau_{ATE,t}^{j}\right]$$

  $\tau_{ATE,t}$ describes how average outcomes would change if GBI were extended to the full study population described in section III.3.a.

- **Average treatment effect among treatment responders (*ATE-R*)**, defined analogously but for the population that responded to follow-up surveys and verified eligibility with the city:

$$\tau_{ATE-R,t} \equiv E\left[Y_{it}(1)-Y_{it}(0) \vee R_{it}(1)=1\right]$$

$$\text{¿}\sum_{j=1}^{8} n_{jt}^{R} E\left[Y_{it}(1)-Y_{it}(0) \vee G_i = j, R_{it}(1)=1\right]$$

$$\text{¿}\, E_t\left[\tau_{ATE-R,t}^{j}\right]$$

[‡‡‡] See Rubin (2005) for one discussion of the potential outcomes framework.

[§§§] An important assumption that we make throughout out study is that potential outcomes are not a function of anyone else's treatment status or outcomes. This is called the Stable Unit Treatment Value (SUTVA) assumption. It is already implicit in the fact that we write potential outcomes only as a function of unit i's *own* treatment status, $Y_{it}(D_i)$. Many factors influence each household's outcomes. Too many, in fact, for us to list. SUTVA states that treated and control units themselves do not affect each other's outcomes.
Whether such spillovers exist and how big they are is an active area of economic research and an important question about how a community-wide GBI would operate (see Daruich and Fernandez 2022). In the context of our study, it is unlikely that GBI receipt by the (relatively small) number of participants will affect outcomes for the much larger group of nonparticipants. However, because our study is small and compares individuals who receive GBI payments to those who do not, we cannot answer questions about how a GBI program implemented at scale would affect outcomes.

[****] In the context of GBI, many other types of causal effects may be of interest. Stockton's SEED demonstration, for example, studies GBI's effect on the *variance* of household income as opposed to the mean. Relatedly, GBI may only affect outcomes for those with the most extreme baseline values. Distributional (or quantile) treatment effects answer questions like that (see Bitler, Gelbach, and Hoynes 2006). GBI may have heterogeneous effects by recipient's characteristics such as gender, race, age, education, or work history. Ideally, a study is designed to estimate this heterogeneity from the outset, but these causal questions require large sample sizes. Because of the limited scale of the Minneapolis GBI pilot, we chose not to target causal parameters related to heterogeneity along these (and other) margins and we do not plan to test for them in our sample.

> $\tau_{ATE-R,t}$ describes how average outcomes would change among the types of treatment households that responded to our surveys if GBI were extended to the full study population. This may differ from the average effect of such a policy if survey respondents react differently to GBI than attritors.

These definitions explicitly show how summary target parameters relate to stratum-specific treatment effects, which are the building blocks in our stratified design. Note that $\tau_{ATE-R,t}$ may change across follow-up waves if either stratum-specific *ATE*s change ($\tau^{j}_{ATE-R,t}$) or if non-response changes the stratum distribution (n^{R}_{jt}). We discuss the weighting of stratum-specific estimates below.

d. Identifying Assumptions

The fundamental challenge to learning about the causal effects of GBI—or a treatment effect in any other study—is that we only observe one of the two potential outcomes for any individual. For those randomized to receive treatment, our survey measurements represent $Y_{it}(1)$, their outcome "under treatment." For those randomized into the control group, we observe $Y_{it}(0)$, outcomes "under control." The causal parameters defined above, however, require an estimate of $E[Y_{it}(0)]$ or $E[Y_{it}(0) \vee R_{it}(1)=1]$, averages of untreated potential outcomes for groups of units that extend beyond the actual control group.

The randomized treatment ensures that data from the full control group reflect these means. Since control group members are a random draw from the study sample, their outcomes show what would have happened had the entire sample been untreated: $E[Y_{it}(0) \vee D_i=0]=E[Y_{it}(0)]$. For the same reason, the average outcome in the randomized treatment group is the same as the average treated potential outcome in the whole sample: $E[Y_{it}(1) \vee D_i=1]=E[Y_{it}(1)]$. Therefore, randomization means that the average outcome for treated units minus the average outcome for untreated units equals the average treatment effect: $E[Y_{it}(1) \vee D_i=1]-E[Y_{it}(0) \vee D_i=0]=E[Y_{it}(1)]-E[Y_{it}(0)]=E[Y_{it}(1)-Y_{it}(0)]=\tau_{ATE,t}$.

Unfortunately, there was substantial attrition. At 6 months 84 percent of the treatment group provided usable data but only 36 percent of the control group did so. The **respondent sample** is not the same as the **full sample** from which we randomly assigned GBI treatment status. Therefore, the mean outcome among control *respondents* might no longer equal the mean of $Y_{it}(0)$ in the full sample, and the mean outcome among treatment respondents might not equal the mean of $Y_{it}(1)$ in the full sample either. The difference of average outcomes for treated and control *respondents* is no longer guaranteed to equal $\tau_{ATE,t}$.

Because of attrition, we need to make additional statistical assumptions to justify treatment/control comparisons generally, or treatment/control comparisons for similar subsets of respondents. As such, our analytical approach more closely follows methods for observational studies, although our design is helped by the initial randomization. In some cases we can test implications of our identifying assumptions and generate evidence to support a specific causal interpretation of our findings.

Here we define the assumptions that allow us to identify stratum-specific parameters and therefore summary parameters as well. The following assumption from Ghanem et al. (2022) formalizes the notion that attrition does not introduce bias *and* preserves the interpretation of such comparisons as $\tau_{ATE,t}$:

Assumption 1. Unconditional Mean Internal Validity for the Population (Mean IV-P)

$$E[Y_{it}(d) \vee S_i=1, G_i, D_i, R_{it}]=E[Y_{it}(d) \vee S_i=1, G_i], d=0,1, t=0,1$$

This assumption says that mean potential outcomes ($Y_{it}(1)$ and $Y_{it}(0)$) for all four combinations of treated/respondent groups are equal in each stratum at baseline and follow-up. It is an implication of a stronger assumption, which is that attrition is random conditional on treatment status. In other words, the respondents we are missing are a random subset of the treatment or control groups.

Under Assumption 1, a simple difference of means (SDM) between treated and control respondents identifies $\tau_{ATE,t}^{j}$:

$$E[Y|S=1,G=j,D=1,R=1]-E[Y|S=1,G=j,D=0,R=1]$$

$$\text{¿}E[Y(1)|S=1,G=j,D=1,R=1]-E[Y(0)|S=1,G=j,D=0,R=1]$$

$$\text{¿}E[Y(1)|S=1,G=j]-E[Y(0)|S=1,G=j]$$

$$\text{¿}E[Y(1)-Y(0)|S=1,G=j]$$

$$\text{¿}E[Y(1)-Y(0)\vee G=j]$$

$$\text{¿}\tau_{ATE,t}^{j}$$

Where the second line substitutes in potential outcomes, the third line uses assumption 1, the fourth line combines terms, and the fifth line follows from the fact that S is a random sample from the population.

A slightly different assumption (also from Ghanem et al. (2022)) formalizes the notion that attrition does not introduce bias, but only permits identification of a mean parameter that is valid for respondents:

Assumption 2. Unconditional Mean Internal Validity for Respondents (Mean IV-R)

$$E\left[Y_{it}(0)\vee S_i=1,G_i,D_i,R_{it}\right]=E\left[Y_{it}(0)\vee S_i=1,G_i,R_{it}\right],t=0,1$$

This assumption says that mean untreated potential outcomes are the same for treated and untreated respondents and are also the same for treated and untreated attritors. It is an implication of a stronger assumption, which is that treatment is random conditional on response status. Therefore, comparing treated to untreated respondents identifies a causal effect. Because Assumption 2 does not put any restrictions on $Y_{it}(1)$, however, it leaves open the possibility that treatment effects are different for respondents and non-respondents, and so we can only use it to identify average treatment effects for respondents.

The following assumption allows us to identify $\tau_{ATE-R,t}$ by a SDM:

$$E[Y|S=1,G=j,D=1,R=1]-E[Y|S=1,G=j,D=0,R=1]$$

$$\text{¿}E[Y(1)|S=1,G=j,D=1,R(1)=1]-E[Y(0)|S=1,G=j,D=0,R(0)=1]$$

$$\text{¿}E[Y(1)|S=1,G=j,R(1)=1]-E[Y(0)|S=1,G=j,R(0)=1]$$

$$¿E\left[Y(1)\middle|S=1,G=j,R(1)=1\right]-E\left[Y(0)\middle|S=1,G=j,R(1)=1\right]$$

$$¿E\left[Y(1)-Y(0)\middle|S=1,G=j,R(1)=1\right]$$

$$¿E\left[Y(1)-Y(0)\vee G=j,R(1)=1\right]$$

$$¿\tau_{ATE-R,t}^{j}$$

Where the second line substitutes potential outcomes for Y and for R, the third line uses assumption 2 (mean independence of Y from D conditional on R), the fourth line uses assumption 2 again (because it implies that mean $Y(0)$ among control-responders is the same as the mean of $Y(0)$ among treatment responders) [††††], the fifth line combines terms, and the sixth line uses the fact that we selected a random sample from the population of interest. It is important to emphasize that these assumptions are strong enough to allow SDM (within strata) to identify causal effects.

We also define conditional versions of these assumptions that allow attrition to be correlated with observed variables.

By conditioning on those variables, using various strategies discussed below, we can then identify $\tau_{ATE,t}$ or $\tau_{ATE-R,t}$ even under certain forms of selective attrition that would bias the SDM estimator.

Assumption 1X. Conditional Mean Internal Validity for the Population (C-Mean IV-P)

$$E\left[Y_{it}(0)\vee S_i=1,G_i,D_i,R_{it},X_i\right]=E\left[Y_{it}(0)\vee S_i=1,G_i,X_i\right]$$

Assumption 2X. Conditional Mean Internal Validity for Respondents (C-Mean IV-R)

$$E\left[Y_{it}(0)\vee S_i=1,G_i,D_i,R_{it},X_i\right]=E\left[Y_{it}(0)\vee S_i=1,G_i,R_{it},X_i\right]$$

Assumptions 1X and 2X work the same way as assumptions 1 and 2, but apply to units with a given value of a set of covariates, X_i. (X_i may contain several covariates so that $X_i=x$ refers to units with specific values of each variable.)

[††††] Note that by conditioning on actual response status, Assumption 2 places restrictions on the relationship between $Y(0)$ and potential outcomes for response status, $R(1)$ and $R(0)$. For example, Assumption 2 implies:

$$E\left[Y(0)\middle|G,R=1,D=1\right]=E\left[Y(0)\middle|G,R=1,D=0\right]E\left[Y(0)\middle|G,R(1)=1\right]=E\left[Y(0)\middle|G,R(0)=1\right]$$

The mean untreated potential outcome among treatment-responders is the same as it is among control responders (ie. that any differential attrition is random with respect to $Y(0)$). Denoting the share of treatment responders induced to respond because of treatment by $\pi\equiv\frac{P(R(1)>R(0)\vee D=1)}{P(R(1)=1\vee D=1)}$, this further implies:

$$\pi E\left[Y(0)\middle|G,R(1)>R(0)\right]+(1-\pi)E\left[Y(0)\middle|G,R(0)=1\right]=E\left[Y(0)\middle|G,R(0)=1\right]$$
$$E\left[Y(0)\middle|G,R(1)>R(0)\right]=E\left[Y(0)\middle|G,R(0)=1\right]$$

This says that the mean untreated potential outcome among “response compliers” is the same as among control respondents.

These assumptions are weaker than assumptions 1 and 2 because they allow for the possibility that attrition creates a respondent sample in which the distribution of determinants of $Y_{it}(0)$, measured by X_i, differ in the treatment and control groups. The most likely way this would occur in our study is if a GBI offer caused very different response rates for different types of participants: $E\left[R_i(1)-R_i(0)\middle|G_i, X_i=x_0\right]\neq E\left[R_i(1)-R_i(0)\middle|G_i, X_i=x_1\right].$ As long as $E\left[Y_{it}(0)\middle|G_i, X_i=x_0\right]\neq E\left[Y_{it}(0)\middle|G_i, X_i=x_1\right]$, then the compositional differences generate by differential attrition mean that assumptions 1 and 2 do not hold.

Suppose, for example, that participants with young children have a harder time filling out our survey instrument than households with school-age children, but that a GBI offer makes it worthwhile for them to do so. In that case, our sample of control respondents will have a smaller share of parents of young children than the treatment respondents. Mothers (mostly) reduce their labor supply when children are young, so this compositional difference between the treatment and control respondents will lead to lower employment rates in the treatment group than the control group even if GBI had no causal effect on employment.

Assumptions 1X and 2X allow us to address this issue by assuming that participants with a given set of characteristics who respond when they are offered treatment have the same average $Y(0)$ as similar participants who respond when not offered treatment (see footnote 13). Assumption 1X further requires average treated potential outcomes to be the same for these groups (ie. participants did stay in the study if offered treatment *because* they anticipate large benefits of treatment). Under these assumptions, if we make our treatment/control comparisons separately for households with younger and older children, then we are able to identify conditional average treatment effects:

$$\tau^{j}_{CATE,t}(x)\equiv E\left[Y_i(1)-Y_i(0)\vee G_i, X_i=x\right]$$

$$\tau^{j}_{CATE-R,t}(x)\equiv E\left[Y_i(1)-Y_i(0)\vee G_i, R_i(1)=1, X_i=x\right]$$

Estimates of $\tau^{j}_{CATE,t}(X_i)$ or $\tau^{j}_{CATE-R,t}(X_i)$ can then be aggregated using the distribution of X_i in the full sample or respondent sample to obtain the overall ATE parameters.

Below we use two types of covariates to identify treatment effects under assumptions 1X and 2X. One includes baseline income, education, and age (X_i^{SES}) and the other is the baseline value of the outcome variable (Y_{i0}).

The final type of identifying assumption that we consider is one that restricts the average change in untreated potential outcomes from baseline to follow-up to be the same in the treatment and control groups:

Assumption PT. Parallel Trends

$$E\left[Y_{it}(0)-Y_{i0}(0)\middle|D_i=1, G_i=j, R_i=1\right]=E\left[Y_{it}(0)-Y_{i0}(0)\vee D_i=0, G_i=j, R_i=1\right]$$

The parallel trends (PT) assumption allows us to identify $\tau_{ATE-R,t}$ parameters using a difference-in-differences (DiD) estimator that compares the average change in outcomes for treated respondents to the average change in outcomes for control respondents. Therefore, if differential attrition skews the mean $Y_{it}(0)$ in the same way in all waves, then a DiD approach accounts for it because outcome changes net out that (constant) bias.

3. Validating the Experimental Design

a. Validating the Randomization

Because we have baseline data on all participants we can test whether our stratified randomization successfully balances the characteristics of treatment and control participants in the full study sample. We focus these **balance tests** on two types of variables: baseline values of our outcome indices and fixed demographic and socioeconomic variables that describe the study population. Indices are defined above. The other variables will include:

- **i.** Education
  - Less than high school (EDUC_SCHCOMP==“Less 8th grade” | “Less than a high school diploma”)
  - High school graduate (EDUC_SCHCOMP==“High school diploma or equivalent”)
  - Some college (EDUC_SCHCOMP==“Some college but no degree”)
  - Post-high school degree (EDUC_SCHCOMP==“Associate’s degree” | “Professional school degree” | “Bachelor’s degree” | “Master’s degree” | “Doctorate degree”)
- **ii.** Gender
  - DEMO_GENDER==“Man”
  - DEMO_GENDER==“Woman”
  - Neither (DEMO_GENDER==“Nonbinary” | Don’t know or prefer not to answer”)
- **iii.** Age
  - DEMO_AGE
- **iv.** Household Size
  - HH_COUNT (top-coded at 11)
- **v.** Distribution of Children
  - Any children
  - Number of children
  - Number of children under age 5
- **vi.** Cumulative Income Distribution
  - Share with income below 13 cutoff values (FIN_HOUSEINC: $5,000, $7,500, $10,000, $12,500, $15,000, $20,000, $25,000, $30,000, $35,000, $40,000, $50,000, $75,000)

We will report a table of means ($\overline{x}_0^T$ and $\overline{x}_0^C$) and standard deviations (s_T^2 and s_C^2) of each variable in the treatment and control group.

Table P2. Balance Test in the Full Study Sample

Outcome	Control	Treatment	Difference	$\sqrt{\frac{s_T^2+s_C^2}{2}}$
Index 1	Mean (s.d.)	Mean (s.d.)		
Index 2				
[…]				
x1				
x2				
[…]				

To make it easier to compare across variables with different units (i.e., some questions have yes/no answers, other refer to numbers of people [living in your household for example], and other answers are dollar amounts), we will report a Love plot of standardized mean differences between treatment and control means:

$$\frac{\bar{x}_0^T - \bar{x}_0^C}{\sqrt{\frac{s_T^2+s_C^2}{2}}}$$

Because our randomization was stratified, a formal test of the randomization needs to condition on strata dummies (i.e., only compare baseline means between treated and control observations in the same stratum). We will jointly estimate mean differences, conditional on strata dummies, across all baseline variables using the randomization-based omnibus test in Hansen and Bowers (2008). This test calculates weighted averages of stratum specific treatment/control differences in each covariate, then forms a joint test statistic by aggregating squared differences by the inverse covariance matrix of the set of imbalance statistics. The covariance matrix is calculated from a series of random permutations of the treatment vector (conditional on the stratification), as is the finite-sample distribution of the test-statistic. This approach allows a single test of the null hypothesis that baseline means (as well as means of linear combinations of our chosen covariates) are equal within strata for all variables.

b. Attrition

Only some of the 530 original sample members chose to continue in the study. Some original participants moved out of the city of Minneapolis. Some participants did not return the necessary documents to verify their income eligibility (stipulated by ARPA). Others exited the study or failed to respond to surveys sent to them. At least one participant passed away between the 6-month and 12-month survey. When participants do not provide usable data either because the leave the study entirely or do not answer all or part of a survey, this is called **attrition**.

The balance test in Table P2 serves as a check on the randomization procedure itself because it uses all 530 randomized households. It cannot not check whether the GBI treatment is still random in the respondent sample. This section describes how we measure attrition, quantify its effect on the characteristics of treatment and control *respondents,* and test whether assumptions 1, 2, 1X, and 2X are likely to hold.

i. Quantifying Attrition

We will produce a table of response rates starting from 100% at baseline and tracing out the share of the sample who respond to a given wave (survey non-response) or respond enough to construct a given index (item non-response).

Table P3A. Attrition Rates by Outcome Domain, Wave, and Treatment Assignment

		Response Rate		
	Wave	**Control**	**Treatment**	**Differential Attrition Test**
A. Index 1				
	6-month	%	%	Coef (s.e., N)
	12-month			
	18-month			
	24-month			
B. Index 2				
	6-month			
	12-month			
	18-month			
	24-month			
[...]				

We will also report a table that describes the number and share of sample members who attrit because they fail to complete eligibility verification or who are enrolled and eligible but do not fill out the survey (details on survey non-response).

Table P3B. Response and Eligibility Verification Counts by Treatment Assignment

	Treatment	Control
Verified and responded		

Did not verify:

moved

income ineligible

did not submit paperwork

Did not respond to survey

Response patterns to the 6-month survey revealed that some control group members filled out surveys but failed to verify their financial or geographic eligibility for the study or to provide financial enrollment paperwork (necessary to receive their $150 survey payment). We do not currently include these participants in the respondent sample. The City will continue to enroll participants, however, throughout the study. Therefore, if these respondents verify their eligibility at a later date, we can add their data to the 6-month respondent sample and re-analyze.

If attrition in one or more strata is so severe that no control members remain, we can no longer estimate average treatment effects in that stratum and will exclude it entirely from analyses of that wave. If that happens, we will report estimates from earlier waves without that stratum to improve comparability of our findings over time.

ii. Characteristics of Respondents and Attritors (Selectivity)

We use the baseline data, which we have for all participants, to measure two important dimensions of attrition. First, we measure differences in the characteristics of attritors and respondents. This helps us interpret our findings. If treatment attritors, control attritors, treatment respondents, and control respondents are all similar, we feel confident that our estimates describe the causal effect of GBI on our target population. The main characteristics we measure in this analysis are baseline outcome values. This is part of the statistical content in Assumptions 1 and 1X.

Second, we measure differences in the characteristics of treated and control *respondents.* Randomization ensures that treatment and control group members have identical potential outcomes (in expectation) in the full sample, but attrition can introduce differences in characteristics of respondents. In other words, a particular type of respondent might stay in the treatment group but exit the control group, leading to a treatment that is not random conditional on response status. A warning sign of this problem would be if treated respondents had very different characteristics from control respondents (and also if treated/control attritors were very different from each other). If treatment and control respondents are similar attritors and responders differ, then our findings describe a causal effect of GBI but for a subpopulation who are willing and able to participate in an ongoing study. This reflects the statistical content of Assumptions 2 and 2X.

Unconditional Tests:

We will implement the tests of assumptions 1 and 2 proposed by Ghanem et al. (2022). We will report a table that contains mean outcome values at baseline for control respondents, control attritors, treatment respondents, and treatment attritors. There may be slightly different samples for each outcome because of item non-response. We will then report *p*-values from a joint test that baseline means are equal by treatment status conditional on strata and response status (a test of assumption IV-R) and a joint test that baseline means are equal across all four groups (a test of assumption IV-P). These come from the following regression specification:

$$Y_{ki0}=\sum_{j=1}^{8} g_{j(i)}\left[\gamma_{j1}^{C}(1-D_i)R_i+\gamma_{j0}^{C}(1-D_i)(1-R_i)+\gamma_{j1}^{T}D_iR_i+\gamma_{j0}^{T}D_i(1-R_i)\right]+\epsilon_{ki0}\,(MIV)$$

Where the null hypothesis to test unconditional IV-R is a joint test of 16 equalities:

$$H_0^{IV-R}:\gamma_{j1}^{C}=\gamma_{j1}^{T}\wedge\gamma_{j0}^{C}=\gamma_{j0}^{T}\,\forall\,j$$

And the null hypothesis to test unconditional IV-P is a joint test of 24 equalities:

$$H_0^{IV-P}:\gamma_{j1}^{C}=\gamma_{j1}^{T}=\gamma_{j0}^{C}=\gamma_{j0}^{T}\,\forall\,j$$

Tests conditional on X_i^{SES}:

One of the specifications we outline below uses baseline age, income, and a higher-education indicator to try and address selective attrition. To test whether this approach is likely to address bias from selective attrition we will also estimate a conditional version of this test by including strata/treatment/X_i^{SES} interactions in the above specification.[‡‡‡‡]

If we reject the null hypothesis for IV-R using the unconditional test, it means that baseline outcomes differ meaningfully between treatment and control respondents in at least one stratum. Simply comparing treatment and control means is thus unlikely to have any causal interpretation. If we reject the null hypothesis for IV-P, we conclude that SDM is unlikely to estimate $\tau_{ATE,t}$, but may still be interpretable as $\tau_{ATE-R,t}$. The degree of confidence we have in the design after accounting for attrition therefore requires both tests. We draw the same conclusions from the conditional tests, but they apply to a SES-adjusted estimator (specification 1 below) instead of SDM.

These conditional tests may show that baseline outcomes differ by treatment status among respondents even after conditioning our comparison (linearly) on X_i. As an alternative, we can directly use data on baseline outcomes to adjust our

[‡‡‡‡] This approach only tests for imbalance in baseline outcome means given the linear specification of equation MIV. Assumptions 1X and 2X are non-parametric assumptions in that they condition on a particular value of X_i without specifying a relationship between outcomes and covariates. Therefore, tests of the joint null hypotheses listed above—based on coefficients from a regression with controls—may be misleading if the regression is misspecified.

comparisons. The estimation section proposes two ways to do this, each of which relies on slightly different identifying assumptions. These assumptions (or their implications) are untestable with only one pre-treatment period.

One drawback of these tests is that they may have low power to detect meaningful differences in baseline mean outcomes because they test many null hypotheses with only 530 observations. To explore a higher-powered test, we will also estimate versions of this test that ignore the strata dummies. Because we chose the same treatment probability in each stratum, the means from this approach will still be valid.

	Baseline Means (Y-bar 0)							
Wave	Control Respondents	Control Attritors	Treatment Respondents	Treatment Attritors	Unconditional IV-P test (Assumption 1)	Unconditional IV-R test (Assumption 2)	Conditional IV-P Test (Assumption 1X)	Conditional IV-R Test (Assumption 2X)
A. Index 1								
6-month	#	#	#	#	p-value	p-value	p-value	p-value
12-month								
18-month								
24-month								
B. Index 2								
6-month								
12-month								
18-month								
24-month								
[…]								

Table P4. Selective Attrition Tests at Baseline

iii. Stratum Sizes in the Respondent Sample

The definition of our parameters of interest in section III.3.c expressed them as averages across stratum-specific average treatment effects. The way these averages are calculated therefore affects our results. Below we proposed to estimate stratum-specific coefficients and weight them together after estimation using fixed weights defined by the baseline stratum size distribution, $n_j \equiv \frac{N_j}{530}$. If Assumption 1 or 1X hold, this approach identifies $\tau_{ATE,t}$ because each stratum-specific estimate equals $\tau^j_{ATE,t}$ and $\tau_{ATE,t}$ is the average of the 8 $\tau^j_{ATE,t}$ parameters weighted by n_j. If Assumption 2 or 2X holds, however, and the stratum size distribution among respondents is different from n_j, then weighting each $\tau^j_{ATE-R,t}$ by n_j instead of $n^R_{jt} = \frac{N^R_{jt}}{\sum_j N^R_{jt}}$, the share of respondents at time t in stratum j, does not equal $\tau_{ATE-R,t}$. Instead, weighting by n_j estimates $\tilde{\tau}_t \equiv \sum_j \tau^j_{ATE-R,t} n_j = \tau_{ATE-R,t} + \sum_j \left(n_j - n^R_{jt}\right) \tau^j_{ATE-R,t}$. One reason to report $\tilde{\tau}_t$ is that it will only change across waves if the within-stratum treatment effect estimates change. On the other hand, even if each $\tau^j_{ATE-R,t}$ is constant over time, $\tau_{ATE-R,t}$ can change across waves if there are different attrition rates across the strata.[§§§§]

We will report a table of the number of observations in each stratum at each wave (reported for the baseline sample in Figure P3) to provide information on different ways to weight the stratum-specific effects.

[§§§§] As described in section 3, we use OLS to estimate stratum-specific effects and then aggregate them using the baseline stratum distribution using post-estimation commands. A more common estimation approach is to report the coefficient on a treatment dummy from an OLS regression with stratum fixed effects (and possibly other control variables). This yields an average of stratum-specific effects weighted by n^R_{jt} and by the within-stratum variance of treatment (Gibbons, Urbancic, Suárez-Serrato 2018). When treatment rates are the same across strata—as they essentially are in our full sample (36-38%)—the variance component cancels and the weighting scheme is based only on strata sizes. But if attrition creates different realized treatment rates across strata, then OLS will put extra weight on treatment/control contrasts from strata with treatment probabilities closer to 0.5 and less weight on contrasts from strata with treatment probabilities approaching 0 or 1. Our approach avoids this issue.

Table P5. Distribution of Respondents Across Strata and Stratum Treatment Probabilities

Strata	Kids	ZIP	Poverty	Respondents (Treatment Probability)			
				Baseline	6 months	12 months	24 months
1				# (%)			
2				# (%)			
3				# (%)			
4				# (%)			
5				# (%)			
6				# (%)			
7				# (%)			
8				# (%)			

If the strata-size distribution changes meaningfully over time we will also calculate treatment effects weighted by n_{jt}^{R} to ensure that (under the necessary assumptions) our results equal $\tau_{ATE-R,t}$.

c. Quantifying the Size of the GBI Treatment

Treated households all received $500 per month, but varied in their baseline income and number of household members. Therefore, GBI means different things to different households. For example, it may have large effects for lower-income households, but smaller effects among households with higher income for whom GBI is a smaller proportional increase in income.

We compare the annual GBI amount of $6,000 to both baseline income and household size to quantify how large the treatment is relative to observed household characteristics.

i. Our baseline survey instrument solicited information on annual household income in 13 bins. For households reporting annual 2021 income between $A and $B, we know that $6,000 represents between an a% = $6,000/$A and b% = $6,000/$B increase in income relative to their baseline. Therefore, we can measure the share of treated households for whom GBI represents different approximate percentage increases over baseline.

   We will present this information in a histogram that plots the share of the treatment group whose baseline income response implies a given range of percentage increases from GBI:

Figure P4. Distribution of the Percent Increase Over Annual Baseline Income from Annual GBI Payments in the Treatment Group

[HISTOGRAM]

We can also use cumulative versions of these statistics to estimate the average percent of baseline income that GBI represents. We will report this statistic as a note in Figure P4.

ii. We also collected data on household size so we can calculate the per-person GBI amount by dividing $6,000 by the number of household members. We will present histogram that plots the distribution of this ratio (e.g., $6,000 for one-person households, $3,000 for two-person households, etc.).

Figure P5. Distribution of Per-Person GBI Payment in the Treatment Group

[HISTOGRAM]

We will also calculate the average per-person GBI amount in the treatment group and report it as a note in Figure P5.

This information contextualizes the treatment effects we estimate.

4. Estimation and Inference

Three key features of the design and the implementation of the pilot guide our estimation approach:

- **Stratification.** Because randomization occurred within each stratum we chose to estimate stratum-specific treatment effects and report explicitly weighted averages of them.
- **Attrition.** Because of substantial and differential attrition at 6 months, we prefer specifications that condition on variables we believe are correlated with attrition and untreated potential outcomes.
- **Baseline data.** Because we have comparable data at baseline for all respondents, some of our specifications use pre-treatment outcomes to control for any bias introduced by differential attrition.

Based on these considerations we chose four specifications that identify the average treatment effect parameters defined in section III.3.c under different assumptions. Finally, we outline two approaches to statistical inference. One strictly controls the probability of any false discovery. We apply this method to an analysis of the indices defined in section III.2.a. The other controls the number of false discoveries. We apply this method to an exploratory analysis III.3.c.

a. Empirical Specifications

Our specifications estimate stratum-specific treatment effects and we use post-estimation commands to construct estimates of the summary parameters defined above:

$$\hat{\tau} = \sum_{j=1}^{8} n_j \hat{\tau}^j$$

Note that the difference between assumptions 1 and 2 (or 1X and 2X) do not alter the estimator—it is either SDM or some kind of conditional comparison within strata—but it does affect how we construct each $\hat{\tau}^j$ and whether we interpret it as $\tau^j_{ATE,t}$ or $\tau^j_{ATE-R,t}$. Note that by choosing a stratified design, all specifications control non-parametrically for the three variables that define strata (poverty, presence of children, and two ZIP code groups) in the sense that we only compare households in the same stratum.

Specification 0 is a simple difference in means within each stratum:

$$Y_{kit} = \sum_{j=1}^{8} g_{j(i)} \left(\mu_j + \hat{\tau}^j_{SDM} D_i \right) + \epsilon_{kit} \left(M0 \right)$$

Where $g_{j(i)}$ are strata fixed effects, $\hat{\mu}_j$ is the mean among control respondents in stratum j and $\hat{\tau}^j_{SDM}$ equals the difference in mean outcomes between treatment and control respondents in stratum j. This estimator identifies $\tau^j_{ATE,t}$ under assumption 1 and $\tau^j_{ATE-R,t}$ under assumption 2.

Specification 1 conditions on pre-treatment demographic and economic covariates and their interactions with strata and treatment dummies.

$$Y_{kit} = \sum_{j=1}^{8} g_{j(i)} \left(\mu_j + \beta_j \dot{X}_i + \rho_j \dot{X}_i D_i + \hat{\delta}^j_X D_i \right) + \epsilon_{kit} \left(M1 \right)$$

In equation (M1), $\dot{X}_i \equiv X_i - \overline{X}_{j(i)0}$ is a vector of covariates demeaned relative to the baseline sample mean in unit i's stratum. With 8 strata and the treatment interaction, each element of X_i uses 16 degrees of freedom. Moreover, with 63% of each stratum assigned to the control group and a 36% response rate among controls, we expect to have 8 or 9 control households in our smallest stratum. Therefore, we cannot include very many covariates and still estimate the $\hat{\delta}^j_X$. We pre-specify income (measured as the mid-point of each respondent's reported category at baseline), a dummy for having a post-high-school degree, and age at baseline as controls in specification 1.

Sloczynski (2022) show that in an interacted specification like (M1), $\hat{\delta}_X^j$ identifies $\tau_{ATE,t}^j$ under assumption 1X because it effectively adjusts the characteristics of *both* the treatment and control groups to reflect the sample-wide (baseline) average $\overline{X}_{j(i)0}$.[*****] Under assumption 2X, however, a consistent estimator of $\tau_{ATE-R,t}^j$ is:

$$\hat{\tau}_X^j \equiv \hat{\delta}_X^j + \rho_j \left(\overline{X}_{j(i)0}^{T,R} - \overline{X}_{j(i)0} \right)$$

where $\overline{X}_{j(i)0}^{T,R} - \overline{X}_{j(i)0}$ measures the mean difference in covariates between treated respondents and the full baseline sample. .[†††††] This is equivalent to the "outcome regression" approach in Heckman, Ichimura, and Todd (1997), in which one regresses outcomes on covariates in the control sample and uses that estimation to predict counterfactual outcomes in the treatment sample. We weight together the constructed $\hat{\tau}_X^j$ parameters from (M1) to obtain our main estimate of τ_{ATE-R}. Because this is identified under the weaker assumption 2X, it is the result that we report in our main findings and on which we conduct statistical inference.

<u>Specification 2</u> conditions on baseline outcomes, $\dot{Y}_{ki0}$, instead of $\dot{X}_{it}$:

$$Y_{kit} = \sum_{j=1}^{8} g_{j(i)} \left(\mu_j + \beta_j \dot{Y}_{ki0} + \rho_j \dot{Y}_{ki0} D_i + \hat{\delta}_{LDV}^j D_i \right) + \epsilon_{kit} (M2)$$

The identifying assumption of this estimator is mean independence between treatment and both potential outcomes (1X) or untreated potential outcomes (2X) conditional on baseline outcomes. Each stratum-specific coefficient, $\hat{\tau}_{LDV}^j$, equals: $\hat{\delta}_{LDV}^j = \left(\overline{Y}_{kt}^{jT,R} - \overline{Y}_{kt}^{jC,R} \right) - \beta_j \left(\overline{\dot{Y}}_{k0}^{jT,R} - \overline{\dot{Y}}_{k0}^{jC,R} \right) - \rho_j \overline{\dot{Y}}_{k0}^{jT,R}$. We construct $\tau_{ATE-R,t}$ estimates as:

$$\hat{\tau}_{LDV}^j = \hat{\delta}_{LDV}^j + \rho_j \overline{\dot{Y}}_{k0}^{jT,R}$$

$$\text{¿} \left(\overline{Y}_{kt}^{jT,R} - \overline{Y}_{kt}^{jC,R} \right) - \beta_j \left(\overline{\dot{Y}}_{k0}^{jT,R} - \overline{\dot{Y}}_{k0}^{jC,R} \right)$$

The time-series relationship between outcomes across periods shapes the way that pre-treatment outcome differences map to estimated counterfactual differences at time t. Note that if pre-treatment outcomes are balanced then $\hat{\tau}_{LDV}^j$ simplifies to the SDM.

[*****] Sloczynski (2022) applies to observational designs. We would not need this result if we had outcome data on the full sample, but given the level of attrition we observe, we motivate our empirical approach using results about non-random observational studies.

[†††††] In the context of a randomized trial with no attrition, Lin (2013) and Negi and Wooldridge (2020) conclude that the variance of this fully interacted estimator in (M1) can be no higher than the variance of the SDM estimator (M0).

Specification 3 is a difference-in-differences (DiD) specification that uses the change in each index from baseline to follow-up as the outcome in a regression with stratum dummies and their interaction with the treatment dummy. It is the same as specification M0 but compares the change in outcomes between baseline and follow-up instead of the level at follow-up:

$$Y_{kit} - Y_{ki0} = \sum_{j=1}^{8} g_{j|i|} \left(\mu_j + \hat{\tau}_{DiD}^{j} D_i \right) + \epsilon_{kit} \, (M3)$$

Each stratum-specific DiD coefficient equals:

$$\hat{\tau}_{DiD}^{j} = \left(\overline{Y}_{kt}^{jT,R} - \overline{Y}_{kt}^{jC,R} \right) - \left(\overline{Y}_{k0}^{jT,R} - \overline{Y}_{k0}^{jC,R} \right)$$

As with specification 2, if pre-treatment outcomes are balanced then $\hat{\tau}_{DiD}^{j} = \hat{\tau}_{SDM}^{j}$, but also if $\beta_j = 1$ then $\hat{\tau}_{LDV}^{j} = \hat{\tau}_{DiD}^{j}$.

An important reason to estimate both specifications 2 and 3 is that as long as either Assumption 2X (with $X_{it} = Y_{ik0}$) or Assumption PT holds, then $\hat{\tau}_{LDV}^{j}$ and $\hat{\tau}_{DiD}^{j}$ fall on either side the relevant treatment effect parameter (they “bracket” $\tau_{ATE-R,t}$, Guryan 2004, Angrist and Pischke 2009, Ding and Li 2019).[‡‡‡‡‡]

[‡‡‡‡‡] The bracketing relationship depends on the sign of the pre-period gap in mean outcomes ($\overline{Y}_{k0}^{jT} - \overline{Y}_{k0}^{jC}$) and whether or not $\beta_j < 1$ (stationarity). Consider a case in which $\beta_j < 1$, which means that the time-series properties of the outcome imply that observed gaps shrink over time, and assumption 2X holds with $X_{it} = Y_{ik0}$. If mean baseline outcomes are higher in the treatment group than in the control group by one standard deviation ($\overline{Y}_{k0}^{jT} - \overline{Y}_{k0}^{jC} = 1$), then we expect the mean gap in untreated potential outcomes in period t to equal β_j. The DiD estimator, however, subtracts the observed pre-period difference of 1. This adjustment is too large (because $1 > \beta_j$) which means that $\hat{\tau}_{DiD}^{j} < \hat{\tau}_{ATE-R,t}^{j}$. On the other hand, suppose that Assumption PT holds so that the gap in untreated potential outcomes that would be observed in period t equals the baseline gap (still assumed to be 1). The LDV specification, on the other hand, subtracts β_j, which is too small so that $\hat{\tau}_{DiD}^{j} > \hat{\tau}_{ATE-R,t}^{j}$.

Our primary results will be reported in the following table:

Table P6. Experimental Results for Index Outcomes

		(1)	(2)	(3)	(4)	(5)	(6)
						Specification	
A. Index 1	Wave	Control Mean (N_C)	Treatment Mean (N_T)	SDM (M0)	X (M1)	LDV (M2)	DiD (M3)
	6-month	# (N)	# (N)	tau-hat	tau-hat (MHT-adjusted p-value)	tau-hat (MHT-adjusted p-value)	tau-hat (MHT-adjusted p-value)
	12-month						
	18-month						
	24-month						
B. Index 2							
	6-month						
	12-month						
	18-month						
	24-month						

Outcomes in the main analysis are: the labor supply index (INDEX_LSUPP), Housing Stability (INDEX_HSTABLE), Financial Security (INDEX_FINSEC), Well-Being (INDEX_WB), Food Security (INDEX_FOODSEC_SECURE), and Psychological Wellness (Kessler 10) (INDEX_K10).

We do not report p-values for the SDM specification for three related reasons. First, early evidence on attrition leads us to believe that post-randomization adjustment will be important. We therefore prefer the point estimates from specifications (M1)-(M3). Second, to the extent that baseline outcomes or covariates predict outcomes, estimates from these specifications will likely be more precise than the SDM estimates, which is another reason to prefer specifications (M1)-(M3). Finally, as discussed below, we report p-values adjusted for multiple hypothesis testing but we do not include the SDM estimates in these adjustments which makes these adjustments less severe. Because we do not adjust their p-values, we do not report them in this table. The SDM point estimates are included for transparency[§§§§§] and to demonstrate how much our statistical adjustments matter.

We will produce a similar table for our exploratory analysis. Outcomes in the exploratory analysis include those listed in section III.1.c.

Table P7. Experimental Results for Exploratory Outcomes

[TABLE, STRUCTURED LIKE P6]

b. Multiple Hypothesis Test (MHT) Adjustments

i. Family-Wise Error Rate Control for Main Analysis

For each of our 3 primary specifications, our main results test 6 hypotheses. If GBI has no effect on any outcome and we reject a given hypothesis at the 5% significance level then this implies an 26.5 percent chance of at least one false rejection (if all null hypotheses are independent): 1-P(all tests fail to reject) = 1-0.95^6= 1-0.735. For each specification in our primary analysis we report p-values that are adjusted to control the probability of any false discoveries (the family-wise error rate; FWER) using the free step-down resampling method in Westfall and Young (1993). We outline the steps involved in this method based on the description in Anderson (2008):

Step 1. Estimate all 6 treatment effects using a given specification (M1-M3), collect the analytical *p*-values, and order them from lowest to highest: $p_1 < p_2 < \ldots < p_6$.

Iterate over the following steps B times denoting each iteration by b:

Step 2. Within each stratum, randomly assign a new (false) treatment variable $D_i(b)$ to respondents with the same number of treated and control observations as we observe in the respondent sample.

Step 3. Re-estimate 6 treatment effects using the same specification, weighting, and analytic standard error estimators as the main analysis, but $D_i(b)$ instead of D_i. Calculate *false p-values*: $p_1^{\text{¿}}(b), p_2^{\text{¿}}(b), \ldots, p_6^{\text{¿}}(b)$.

Step 4. Rank these false p-values from lowest to highest and rename them according to the original ordering of significance from step 1: $p_1^{\text{¿}*\text{¿}(b)<p_2^{\text{¿}*\text{¿}(b)<\ldots<p_6^{\text{¿}*\text{¿}(b)\text{¿}}\text{¿}}$. In other words, $p_1^{\text{¿}*\text{¿}(b)\text{¿}}$, the false p-value for test 1, the most significant result in

[§§§§§] A common strategy to gauge how much attrition might affect a study's conclusions is to use the trimming method in Lee (2009). In our case that would involve re-estimating treatment effects after selectively dropping enough treatment observations so that we observed the same share of the treatment and control groups. By dropping the treated units with the highest and lowest outcome values, this approach provides bounds on the causal effect among units who would respond even without treatment. We may implement this approach at the end of our study, but chose not to rely on it as a robustness check because the differences in response rates between treatment and control groups in our study were so large that the bounds would not have been informative. If GBI shifts distribution of a normally distributed without changing its standard deviation, then we can use known features of the normal distribution to calculate how much trimming will change the treatment means. Given our observed differential attrition a bounding estimate would trim about half of the treatment observations. The conditional mean of a normal random variable below zero is -0.8, which implies that trimming the top half of treatment respondents could reduce their observed mean by 4/5ths of a standard deviation. Therefore, if the estimated GBI effect is positive, the lower bound estimated from the trimming approach will be negative unless the point estimate itself is more than about 0.8 standard deviations. Compared to other estimates in the literature on cash transfer programs, this is a large effect (see section 5.b).

from the main estimates, is the minimum of set of false p-values $\{p_1^{\text{¿}}(b), p_2^{\text{¿}}(b), \ldots, p_6^{\text{¿}}(b)\}$. The false p-value for test 2, the second most significant main results, is the minimum of the set of false p-values after removing the lowest one, and so on.

<u>Step 5</u>. For each test, record whether the original p-value, p_r exceeds the simulated p-value $p_r^{\text{¿}*\text{¿}(b)\text{¿}}$ and define a dummy $h_r(b)=1\text{¿}$ that equals one if this condition holds.

<u>Step 6</u>. Calculate the share of iterations (resampled treatment variables) for which $h(b)=1$ and call these *simulated p-values*: $p_r^{fwer*\text{¿}=\frac{\sum_b h_r(b)}{B}\text{¿}}$. Note that $p_r > p_r^{\text{¿}*\text{¿}(b)\text{¿}}$ means that a randomly drawn treatment variable has produced a smaller p-value than the true treatment variable produced, and if this happens often across iterations it shows that random precise findings are common.

<u>Step 7</u>. Rank these simulated p-values, $p_r^{fwer*\text{¿}\text{¿}}$, from lowest to highest and rename them according to the original ordering of significance from step 1: $p_1^{fwer} < p_2^{fwer} < \ldots < p_6^{fwer}$. In other words, p_1^{fwer}, the *adjusted p-value* for test 1, the most significant result in from the main estimates, is the minimum of set of simulated p-values $\{p_1^{\text{¿}}(b), p_2^{\text{¿}}(b), \ldots, p_6^{\text{¿}}(b)\}$. The *adjusted p-value* for test 2, the second most significant main results, is the minimum of the set of simulated p-values after removing the lowest one, and so on. Table P4 will report p^{fwer} values.

ii. False Discovery Rate Control for Exploratory Analysis

When we study exploratory outcomes described in section III.1.c we adjust our p-values to control the share of false discoveries, called the False Discovery Rate (FDR), instead of the probability of *any* false discovery. This introduces a higher risk of incorrectly concluding that GBI has affects a specific outcome, but improves our ability to detect small (true) effects on each exploratory measure, called statistical power.

The basic procedure follows Benjamini and Hochberg (1995) adapted from Anderson (2008):

Step 1. Choose a significance level $q \in (0,1)$. Denote the number of hypotheses being tested by M. In this analysis we are adjusting for multiple hypothesis tests across outcomes within each specification, so M is the number of exploratory outcomes.

Step 2. Rank the M p-values $p_1 < p_2 < \ldots < p_M$ denoting the rank by r.

Step 3. Reject all hypotheses for which $p_r < qr/M$. Note that $r/M \in (0,1)$ is the p-value's percentile rank and q is a factor that further scales this down. With M=10, if the minimum p-value is greater than 0.1 then no hypotheses are rejected.

Step 4. Repeat steps 1-3 for different values of q. The FDR-adjusted p-value for a hypothesis is the lowest q for which it is rejected.

We further sharpen this procedure following Benjamini, Krieger, and Yekutieli (2006). In place of Step 4 do:

Step 5. Set $q^{'}=\frac{q}{1+q}<q$ and let c be the number of hypotheses rejected. This is a more stringent significance level than q itself.

Step 6. Define $\hat{m}_0=M-c$ as the number of potentially true null hypotheses.

Step 7. Repeat steps 1-3 at level $q^{¿}=q^{'}M/\hat{m}_0$.

This approach generates “q-values” that show the significance level at which a given hypothesis can be rejected controlling the FDR.

c. Qualitative Analysis

Our surveys include open-ended questions into which respondents type free text answers. We will use these responses to inform the interpretation of our quantitative findings and consider more formal qualitative evaluation methods if we have resources to pursue this and there are sufficient free-form responses to allow it.

5. Interpretation

While we defer interpretation of our specific findings until after results are released, there are two ways to evaluate plausible treatment effect magnitudes prior to executing our study.

The first is to evaluate what effect sizes our design is likely to be able to detect. This is called **statistical power**: the ability of our sample size, study design, and estimation approach to detect true differences using statistical tests. If we find that our study has low power it will mean that the only treatment effects we could uncover would be extremely large ones. Further research with a stronger research design or more data would then be needed to determine whether GBI had other smaller effects. Power calculations are speculative because they rely on pre-evaluation guesses about key features of our data (such as how well baseline outcomes or covariates predict post-treatment outcomes or how much treatment effects vary).

We also turn to related GBI research to discuss **plausible effect sizes**. Many GBI pilots are underway in the US as of March 2023, but only the SEED demonstration in Stockton, California has released results comparable to what we will estimate. We summarize what has been found in Stockton as a guide to how large the causal effects of Minneapolis’ GBI pilot might be. This discussion is speculative as well because GBI may simply have different effects in the Minneapolis context than in Stockton. In future reports we will also draw comparisons between our findings and non-experimental studies of cash transfers (such as the Alaska Permanent Fund Dividend; Jones and Marinescu [2022]), accounting for differences in the payment amounts, social context, and target groups.[******]

[******] An alternative way to judge the plausibility of a given effect size would be to appeal to a behavioral model, such as a static model of labor supply for example, and use evidence on how low-income households respond to *different* income changes to guess about how GBI respondents would respond in our study. This approach is outside

a. Power Calculations

We calculated power for a range of effect sizes for SDM, LDV, and DiD specifications using the simulation approach in Burlig et al. (2020). We constructed the outcome in two steps.

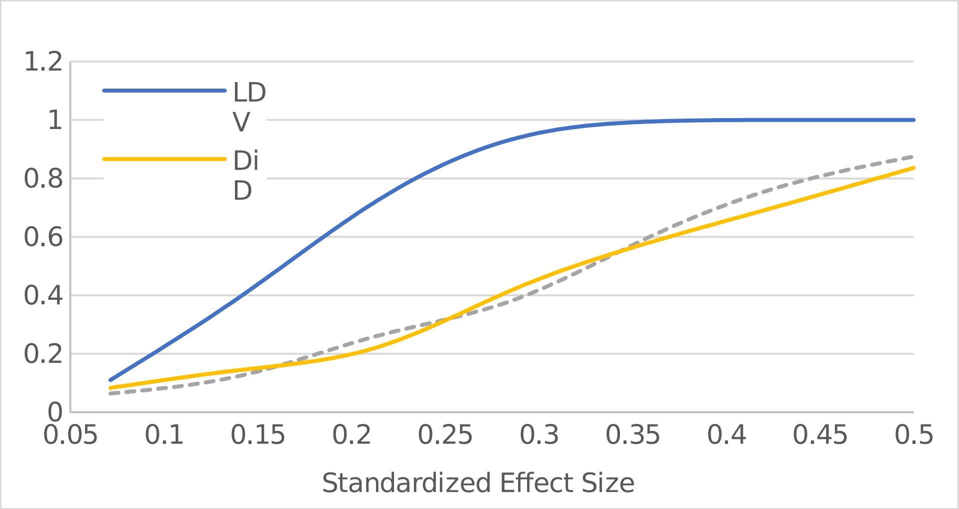
Step 1. Use the variables in INDEX_WB to construct a series of dummies for being in the top two categories of "wellbeing". Create an equally weighted average of these dummies in the baseline sample and keep the 6-month respondent sample only. This variable differs from our main indices in that it does not standardize each component (although all are measured in percentage points) and it uses equal weighting instead of inverse covariance weighting.

Step 2. Create a false post-treatment outcome for each respondent that equals their baseline outcome plus $\frac{v_i}{2} - 0.25$ where $v_i \; U(0,1)$. Censor these generated observations at 0 and 1. This ensures a high degree of persistence in the false outcomes, as we expect in out sample.

We then used the command `pc_simulate` in Stata to calculate power for effect sizes: 0.02, 0.04, 0.06, 0.08, 0.10, 0.12, and 0.14. The standard deviation of our outcome variable at baseline was 0.28, so these effect sizes represent between 7% and 50% of a baseline standard deviation. Our specification incorporated our 8 strata, but assumed an equal treatment probability of 0.375 in each one. This is the treatment allocation in our full sample, but may not hold exactly in our respondent sample. We calculated power for an SDM estimator (M0), a lagged dependent variable estimator (similar to M2, also sometimes called ANCOVA), and for a DiD specification (M3). The results are shown below:

the scope of our pre-analysis plan.

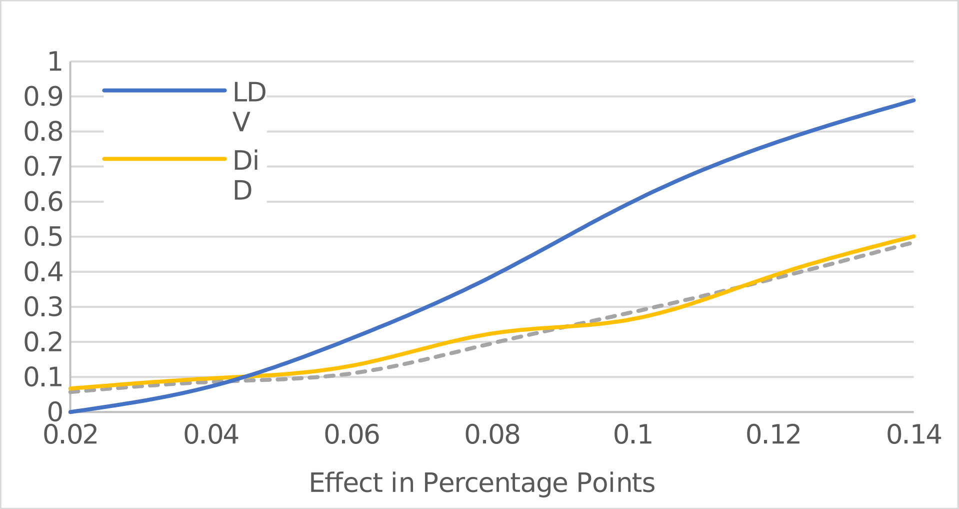
Figure P6. Estimated Power Curves for an Index Outcome



This analysis suggests that our LDV specification has the most power and can detect changes of about one-quarter of a standard deviation with 80% power. The SDM and DiD specifications are similar and achieve 80% power for effect sizes of about 0.45 standard deviations.

In Figure P7 we repeat this power analysis for a dummy variable measuring employment. (We created the follow-up outcome by randomly switching the employment status of 7 percent of the respondent sample from their baseline value.) We plot power curves for this outcome against the percentage point effect on employment. Using this approach our design is not powered to detect reasonably sized employment effects. Only in the LDV specification and only for employment effects of 12 percentage points or more do we achieve 80% power. The baseline employment rate in our study was 53%, so this represents more than a 20% change in employment rates in either direction.

Figure P7. Estimated Power Curves for a Binary Outcome (Employment)



We may come to different conclusions about the power of our study and of different specifications if we constructed our false post-treatment outcomes in a different way. For example, the power of LDV or DiD versus SDM depend strongly on how persistent the outcome variable is (for our constructed index measure the correlation between baseline and follow-up observations is about 0.98, and for the employment measure it is 0.88). This in turn depends on a range of external factors, behavioral factors, and the time elapsed between baseline and follow-up (McKenzie 2012). We do not attempt to capture all of these dimensions in our *ex ante* power analysis. Instead, the conclusion we take from Figure P6 is that the effect sizes that we may be able to detect are not impossibly large and that given the uncertainty around how our outcomes will behave, there is value in estimating a range of specifications. Figure P7, however, suggests that we may not have the statistic power to detect plausible effects on all outcomes of interest, especially not on binary outcomes.

b. Effect Sizes from Related Research

To more tightly benchmark our power analysis, we turn to the closely related findings from the Stockton SEED Demonstration. The Stockton treatment was the same as in our study: $500 per month for 24 months. The program was targeted to neighborhoods in Stockton with low median income (under $46,033 in 2019), but not restricted to households with low incomes. Two reported outcomes include enough information to use as benchmarks for our study: the Kessler 10 scale and current employment. The following table reports pre-treatment and one-year-post-treatment means, estimated treatment effects (using either an SDM or DiD estimator), and standardized effect sizes (treatment effect divided by pre-treatment control standard deviation).

Table P8. Experimental Results from Stockton SEED Demonstration

Outcome	Control Means (s.d.)		Treatment Means (s.d.)		Estimator	Treatment Effect	*Standardized Effect Size*
	Pre	Post	Pre	Post			
Kessler 10	20.7	21.15	21.28	18.43	SDM	-2.72	*-0.30*
	(9.03)	(10.55)	(8.97)	(8.66)	DiD	-3.3	*-0.37*
Current Employment	0.32	0.37	0.28	0.4	SDM	0.03	*0.59*
	(0.05)	(0.05)	(0.05)	(0.05)	DiD	0.07	*1.38*

Effect sizes range from 0.3 to 1.38 standard deviations across outcomes and estimators. At baseline our study population has higher employment rates (53%) and worse mental health scores (25.6 [s.e.=9.8]) than the Stockton sample, and both outcomes have higher standard deviations. It is reassuring that the effect sizes found in Stockton are roughly the same size as the effects that our study is powered to detect.

References

Michael L. Anderson (2008). Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects, Journal of the American Statistical Association, 103:484, 1481-1495, DOI: 10.1198/016214508000000841

Angrist & Jörn-Steffen Pischke, (2009). "Mostly Harmless Econometrics: An Empiricist's Companion," Economics Books, Princeton University Press, edition 1, number 8769.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300. http://www.jstor.org/stable/2346101

Benjamini, Y., Krieger, A.M., & Yekutieli, D. (2006) Adaptive linear step-up procedures that control the false discovery rate. Biometrika. 93(3), 491-507.

Bitler, Marianne, P., Jonah B. Gelbach, and Hilary W. Hoynes (2006). "What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments." *American Economic Review*, 96 (4): 988-1012.DOI: 10.1257/aer.96.4.988

Daruich, Diego and Raquel Fernandez, (2021). “Universal Basic Income: A Dynamic Assessment.” NBER Working Paper 27351

Ding, P., & Li, F. (2019). A Bracketing Relationship between Difference-in-Differences and Lagged-Dependent-Variable Adjustment. Political Analysis, 27(4), 605-615. doi:10.1017/pan.2019.25

Gibbons, Charles E., Suárez Serrato, Juan Carlos and Urbancic, Michael B. (2019). "Broken or Fixed Effects?" *Journal of Econometric Methods*, vol. 8, no. 1, pp. 20170002. https://doi.org/10.1515/jem-2017-0002

Guryan, Jonathan, (2004). "Desegregation and Black Dropout Rates." *American Economic Review*, 94 (4): 919-943.DOI: 10.1257/0002828042002679

Ben B. Hansen, Jake Bowers, (2008). "Covariate Balance in Simple, Stratified and Clustered Comparative Studies," Statistical Science, Statist. Sci. 23(2), 219-236.

Heckman, James J., Hidehiko Ichimura, and Petra Todd, (1997). “Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme.” *The Review of Economic Studies*, vol. 64, no. 4, pp. 605–54. *JSTOR*, https://doi.org/10.2307/2971733. Accessed 2 May 2023.

Jones, Damon, and Ioana Marinescu, (2022). "The Labor Market Impacts of Universal and Permanent Cash Transfers: Evidence from the Alaska Permanent Fund." *American Economic Journal: Economic Policy*, 14 (2): 315-40.DOI: 10.1257/pol.20190299

Lee , David, (2009). Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects, *The Review of Economic Studies*, Volume 76, Issue 3, July 2009, Pages 1071–1102, https://doi.org/10.1111/j.1467-937X.2009.00536.x

Lin, Winston, (2013). "Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique," The Annals of Applied Statistics, Ann. Appl. Stat. 7(1), 295-318.

Negi, Akanksha and Wooldridge, Jeffrey, (2021). Revisiting regression adjustment in experiments with heterogeneous treatment effects, *Econometric Reviews*, 40, issue 5, p. 504-534, https://EconPapers.repec.org/RePEc:taf:emetrv:v:40:y:2021:i:5:p:504-534.

Donald B Rubin (2005). Causal Inference Using Potential Outcomes, Journal of the American Statistical Association, 100:469, 322-
331, DOI: 10.1198/016214504000001880

Tymon Słoczyński (2022) Interpreting OLS Estimands When Treatment Effects Are Heterogeneous: Smaller Groups Get Larger Weights. *The Review of Economics and Statistics* 2022; 104 (3): 501–509. doi: https://doi.org/10.1162/rest_a_00953

West, Stacia, Amy Castro Baker, Sukhi Samra, Erin Coltrera (2021). “Preliminary Analysis: SEED's First Year”

https://static1.squarespace.com/static/6039d612b17d055cac14070f/t/6050294a1212aa40fdaf773a/1615866187890/SEED_Preliminary+Analysis-SEEDs+First+Year_Final+Report_Individual+Pages+.pdf

Peter H. Westfall and S. Stanley Young (1993). Resampling-Based Multiple Testing. Examples and Methods for p-Value Adjustment:: New York: Wiley, ISBN 0-471-55761-7, pp.340