THE MILBANKQUARTERLY

Original Scholarship

Can Social Policies Improve Health? A Systematic Review and Meta-Analysis of 38 Randomized Trials

EMILIE COURTIN ⓘ,* SOOYOUNG KIM ⓘ,[†]
SHANSHAN SONG,[†] WENYA YU,[†]
and PETER MUENNIG ⓘ[†]

Harvard Center for Population and Development Studies, Harvard University; [†]Mailman School of Public Health, Columbia University

Policy Points:

- Social policies might not only improve economic well-being, but also health. Health policy experts have therefore advocated for investments in social policies both to improve population health and potentially reduce health system costs.
- Since the 1960s, a large number of social policies have been experimentally evaluated in the United States. Some of these experiments include health outcomes, providing a unique opportunity to inform evidence-based policymaking.
- Our comprehensive review and meta-analysis of these experiments find suggestive evidence of health benefits associated with investments in early life, income support, and health insurance interventions. However, most studies were underpowered to detect health outcomes.

Context: Insurers and health care providers are investing heavily in nonmedical social interventions in an effort to improve health and potentially reduce health care costs.

Methods: We performed a systematic review and meta-analysis of all known randomized social experiments in the United States that included health

outcomes. We reviewed 5,880 papers, reports, and data sources, ultimately including 61 publications from 38 randomized social experiments. After synthesizing the main findings narratively, we conducted risk of bias analyses, power analyses, and random-effects meta-analyses where possible. Finally, we used multivariate regressions to determine which study characteristics were associated with statistically significant improvements in health outcomes.

Findings: The risk of bias was low in 17 studies, moderate in 11, and high in 33. Of the 451 parameter estimates reported, 77% were underpowered to detect health outcomes. Among adequately powered parameters, 49% demonstrated a significant health improvement, 44% had no effect on health, and 7% were associated with significant worsening of health. In meta-analyses, early life and education interventions were associated with a reduction in smoking (odds ratio [OR] = 0.92, 95% confidence interval [CI] 0.86-0.99). Income maintenance and health insurance interventions were associated with significant improvements in self-rated health (OR = 1.20, 95% CI 1.06-1.36, and OR = 1.38, 95% CI 1.10-1.73, respectively), whereas some welfare-to-work interventions had a negative impact on self-rated health (OR = 0.77, 95% CI 0.66-0.90). Housing and neighborhood trials had no effect on the outcomes included in the meta-analyses. A positive effect of the trial on its primary socioeconomic outcome was associated with higher odds of reporting health improvements. We found evidence of publication bias for studies with null findings.

Conclusions: Early life, income, and health insurance interventions have the potential to improve health. However, many of the included studies were underpowered to detect health effects and were at high or moderate risk of bias. Future social policy experiments should be better designed to measure the association between interventions and health outcomes.

Keywords: social experiments, randomized controlled trials, policy analysis, population health, social determinants of health.

A RISTOTLE WAS PERHAPS THE FIRST TO HYPOTHESIZE THAT one's socioeconomic environment is a key determinant of one's health and longevity.[1] In 1848, the famous Austrian pathologist Rudolf Virchow conducted an epidemiological study showing the strong association between infection with typhus and socioeconomic status.[2] In his report, he recommended social investments as a means to reduce typhoid infections and advocated for democratic institutions to support those investments. Since then, a host of associational studies have solidified strong and consistent relationships between socioeconomic circumstances and an array of measures of physical or mental

health and longevity.[3–5] In the United States, low socioeconomic position is associated with a larger burden of disease than smoking and obesity combined.[6]

These observations from associational studies have led health policy experts around the world to propose policies targeting early childhood development, educational attainment, poverty, housing, and employment as ways to improve population health and reduce health system costs.[7–11] For example, the United Kingdom explicitly mentioned health system savings as a reason to increase social expenditures in the 1990s.[12] Similarly, in the Obama administration, health care cost was the primary motivation behind the creation of incentives for providers to invest in social services for their patients.[13,14] As a result of these incentives, the US Centers for Medicare and Medicaid Services, states, and private providers have together invested billions of dollars in social policies as removed from medical care as high school graduation and housing programs.[13,15] The hope is that these investments will prevent disease in the segment of the US population with the highest health risk—poor and low-income households.

These investments have strong intuitive appeal because they have the potential to improve both the economic well-being and the health of a population. Social policies influence health indirectly, by virtue of their effect on social determinants of health such as education, income, employment or housing. Since these factors are key causes of health, they can in turn affect health outcomes. Finding a way to prevent disease using social policy is no small matter in a nation that has experienced decades-long declines in health and soaring health costs;[16] US health sector costs have swollen to the size of Great Britain's gross domestic product.[17] However, robust evidence suggesting that these programs will be a cost-effective way to improve population health is lacking. A key issue is that the relationships within associational studies can be confounded and may plausibly work in a reverse as well as a forward direction. For example, consider the correlation between income and health. Sick people are more likely to lose their jobs, and as a result, they suffer a loss of incomes.[18] Therefore, it is impossible to determine the extent to which poor health drives down income or low income adversely affects health. Similarly, sick children face challenges in advancing through the education system.[19]

Some investigators have sought to overcome problems associated with confounding and reverse causality by using quasi-experiments to study

the health effects of social policies such as compulsory schooling laws,[20] expansions of the Earned Income Tax Credit,[21] and the distribution of casino earnings on Native American reservations.[22] These types of studies suggest that it is possible to obtain positive health returns on social investments. Quasi-experimental studies are important because they approximate the effectiveness of real-world policies. However, they remain susceptible to residual confounding and other potential biases.[23] For example, state-level policy changes can come from external forces that influence both policymaking and state wealth or health, such as migration, immigration, or fracking.[24–26] Moreover, it may be difficult to publish null or negative findings of quasi-experimental studies because reviewers who believe in the efficacy of social policies might be inclined to reject the results of a well-conducted study with results that challenge their expectations. When only positive findings are published, associations between social policy and health can appear robust when they are not.[27] Randomized controlled trials (RCTs) of social interventions overcome some of these problems. Because participants are randomly allocated, the internal validity of RCTs tends to be high. Therefore, negative findings might be more likely to attract the attention of editors, reviewers, and the media. Because RCTs tend to be conducted by private firms, it is common for all study outcomes to be available, even if they are not always published in peer-reviewed literature.[28]

The United States has a unique history of social experimentation that began in the 1960s.[29] Major, often multi-center, randomized trials were carried out for a wide range of public policies in the 1980s, the "golden age of evaluation."[30,31] Some of these RCTs included health outcomes, providing the opportunity to assess whether changes in socioeconomic conditions induced by the experiments can also affect health. Remarkably, many of these studies have been overlooked in health policy circles. This may be because the major research firms that carried out most of these experiments—Abt Associates, Mathematica, and MDRC—tended to be staffed by economists and sociologists who were less focused on health.[28] Because few health researchers are aware of these large social policy RCTs, these studies are not considered to be part of the canonical literature on the social determinants of health.[32]

To the best of our knowledge, the effects of social experiments on health have not been systematically reviewed. This is remarkable given the magnitude of expenditures in the nonmedical determinants of health in the United States and the weak evidence base supporting these

investments. To date, reviews of the health impacts of social policies have mostly focused on quasi-experimental studies.[33,34] A comprehensive review and meta-analysis that describes and synthesizes the experimental evidence about the impact of socioeconomic interventions on health outcomes is important. Regulators would not allow a blockbuster drug into the health market without extensive experimental evaluation to demonstrate efficacy and safety, but standards for social policy are not as stringent. In our study, we conducted a complete inventory of all known social policy experiments conducted in the United States that included health outcomes. Our intention is to inform the current policy debate on integrating social determinants of health into health care delivery. As such, we emphasize the risk of bias assessments, unearthing null findings and testing for potential false negatives that can arise when interventions are of inadequate duration or have inappropriate sample sizes or follow-up intervals for the outcome that is being measured.

Methods

Search Strategy and Selection Criteria

We conducted our systematic review and meta-analysis according to a protocol registered with PROSPERO (CRD4201810351). We followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) checklist when reporting our findings.[35] This study did not require institutional review board approval.

Our goal was to conduct a comprehensive review of experimental studies of social and economic interventions that were not explicitly designed for the purpose of improving the health of participants. Therefore, studies were eligible for inclusion if they examined a social intervention (eg, income support or housing) evaluated as an RCT (called "social experiment" hereafter) and included health outcomes. Although varied in size and areas of interest, all social experiments considered had to meet the following criteria:[29,36] (a) individuals, families, or larger clusters such as schools or classrooms were randomly assigned to either the intervention or control groups; (b) the treatment group was subjected to a change in socioeconomic incentives or opportunities while the control group was not; and (c) data were obtained on health outcomes. We excluded studies that were not explicitly conducted for the purpose of improving socioeconomic circumstances (eg, interventions that would be reviewed

in the *Community Guide to Preventive Services* or are currently part of the health policy landscape).[37] For example, we excluded the Nurse-Family Partnership program[38] from our review because it was designed as a health intervention. Following Greenberg and colleagues,[29] we also excluded evaluations of social programs that did not use random assignment and trials that did not test social policies (eg, behavioral experiments). We included interventions targeted at US residents, with any subpopulation focus including disabled or aging population programs. We excluded interventions implemented outside of the United States because the policy environments in other countries are different. Online Appendix 1 summarizes our study eligibility criteria.

To identify studies for possible inclusion, we first searched for published studies in the following electronic bibliographic databases: Cochrane Public Health Group Specialized Register, Cochrane Central Register of Controlled Trials, Ovid MEDLINE, Academic Search Premier, Business Source Complete, CINAHL, EconLit, 3IE, PsycINFO, Social Sciences Citation Index, TRoPHI, and WHOLIS. Second, we searched the following grey literature databases: ProQuest Dissertations and Theses, System for Information on Grey Literature in Europe–Open-Grey, EconPapers, National Bureau of Economic Research, the Directory of Open Access Repositories–OpenDOAR, and Social Science Research Network–SSRN eLibrary. Third, we manually searched the websites of companies that have been contracted to conduct these social experiments: MDRC, Abt Associates, Mathematica, and RTI. Fourth, we hand searched reference lists and citations of included studies for additional relevant studies. Finally, we searched the *Digest of Social Experiments* published in 2004 by the Urban Institute.[28] We did not apply any language or time restrictions to our searches. Searches were updated in January 2019. See Online Appendix 2 for an overview of our search strategy.

Two reviewers (Kim and Song) independently assessed all search results by title and abstract, and they independently assessed by full text those studies that were potentially eligible based on our inclusion criteria. Any disagreements were resolved through discussion or arbitration by Courtin.

Data Extraction

Two reviewers (Kim and Courtin) independently extracted data on each study using a customized extraction form. The form was piloted on five

studies before the reviewers proceeded with the full extraction. Study authors were contacted as needed for clarification or in cases of missing data. We extracted data on study authors, publication year, social experiment characteristics (including the study's target population, design, generosity, duration, and follow-up), sample size by group, primary socioeconomic outcome of interest and the experiment's effect on this outcome, health outcomes measured, and number of events in the treated and control groups (for odds ratios).

We first described the studies included in our sample in terms of health outcomes, sample size, population, duration, and follow-up. Next, we categorized social interventions into policy domains based on the established social determinants of health over the life course.[39,40] We covered the following domains: early life and education, income maintenance and supplementation, employment and welfare-to-work, and housing and neighborhood interventions. Because health insurance interventions were primarily designed to provide financial protection for low-income families, we included them as economic interventions.[41]

Within each category, point estimates, corresponding 95% confidence intervals, and nominal statistical significance ($P < .05$, "yes, no") were collected where available. The following steps were followed to select which estimates to report in our narrative review and subsequent analyses: (a) we prioritized models that included the overall study sample over subgroup analyses; (b) we selected the more rigorous models (eg, intent-to-treat estimates over per-protocol analyses); (c) for studies that examined multiple health outcomes, we extracted data for each health outcome; (d) for studies that included repeated measures over time, we extracted data for the latest available time point; (e) if separate models were conducted across different study sites or for children and adults separately, we collected data for each group. We ended up with a sample of 451 unique health estimates across 38 interventions reported in 61 studies.

Risk of Bias Assessment

We used the latest version of Cochrane's risk of bias 2.0 tool to assess the risk of bias in each individual study.[42] Risk of bias was assessed in the following domains: randomization process, deviation from intended intervention, missing outcome data, and selection of reported results. As recommended, studies with high risk of bias in one domain were

classified as high risk of bias overall. Two reviewers (Song and Yu) independently assessed risk of bias, with disagreements resolved through discussion or arbitration by Courtin.

Power Analyses

When reporting estimates of the effects of social experiments on health, false negatives are a potential issue. Typically, a power analysis is conducted a priori on the primary outcomes of interest to determine the number of participants that would be needed to avert a false-negative finding. However, in our review, health outcomes included in the social experiments presented a specific challenge: health measures were typically not included as primary outcome measures, so their statistical power was often not assessed at the design phase of the trial. We therefore conducted post hoc power calculations for each outcome in each study for which relevant parameters were available. For binary outcomes, the power was calculated based on incidence in each intervention group, group sample sizes, and an α-level of 5%. For continuous outcomes, retrospective power calculations were based on available measures of variance (standard deviation [SD] or standard error [SE]), significance level obtained, sample sizes for each group, as well as the number of covariates and R^2 in the case of regression estimates. Key parameters were missing in a number of studies, but we were able to calculate post hoc power for 75% of the estimates included in the review (N = 336 estimates).

Synthesis of Findings

We first provided a narrative description of the studies included in our sample in terms of policy domain, health outcome, sample size, population, duration, and follow-up. We also conducted a qualitative assessment across all outcomes considered in eligible social experiments. For each estimate in our sample, we documented whether the intervention was associated with a statistically significant effect (defined as $P <$ .05) and whether this effect was positive or negative. In the case of fertility outcomes, we categorized reduced fertility as an improvement on the basis of the literature linking lower fertility to higher human capital.[43]

In addition, we conducted a formal meta-analysis on the subsample of studies for which both the category of the experiment (eg, income support) and the outcome (eg, self-rated health) lined up (N = 20

studies). The outcomes that overlapped between experiments were self-rated health (good or excellent self-rated health vs poor), currently smoking (vs not), and obesity (defined as body mass index [BMI]$\geq$30 vs BMI <30) To conduct the meta-analysis, we extracted the number of respondents with the event of interest as well as the total number of respondents in both the treatment and control groups. We then calculated adjusted ORs and associated 95% CIs by combining these numbers in pooled random effects meta-analyses. The random effects model assumes that health effects are heterogeneous and drawn from a normal distribution.[44] Effect heterogeneity was measured using the I^2 statistic.[45]
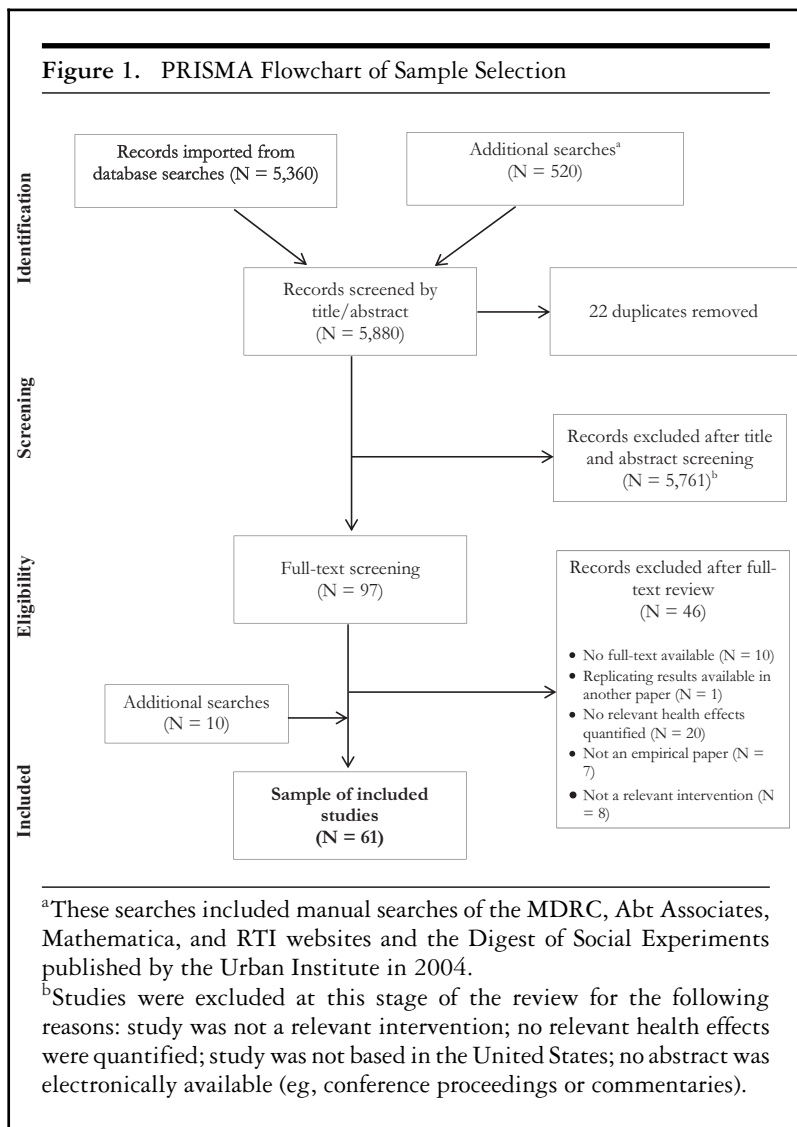
Finally, we carried out analyses to determine which study characteristics were associated with statistically significant findings (vs nonsignificant) and improvement in health (vs worsened health or null effect). These analyses were carried out using multivariate logistic regressions on our sample of 451 primary estimates. We examined the association between study outcomes and publication in a peer-reviewed journal, publication year, sample size, duration of the intervention, duration of follow-up, whether the primary socioeconomic outcome of interest in the trial improved, and whether more than one site was analyzed. To account for correlated results among models presented in the same paper, we clustered the standard errors at the study level. All analyses were conducted in Stata 14.

Results

We identified 5,360 citations from bibliographic databases and an additional 520 from other sources, such as white paper reports from the firms that conducted the experiments. After removal of duplicates, 5,880 unique citations were screened by title and abstract. Of these, we sourced 97 full-text articles for inclusion. Ten additional studies were considered relevant based on manual searching of the reference lists. Sixty-one publications covering 38 social experiments fit our inclusion criteria (Figure 1).

Sample Characteristics

Table 1 lists the main characteristics of our sample of social experiments. Most of the included studies (65%) were published in peer-reviewed journals. Sample sizes ranged from 52 to 74,922 (median = 1,866;

Figure 1. PRISMA Flowchart of Sample Selection

Identification

Records imported from database searches (N = 5,360)

Additional searches[a] (N = 520)

Records screened by title/abstract (N = 5,880)

22 duplicates removed

Screening

Records excluded after title and abstract screening (N = 5,761)[b]

Eligibility

Full-text screening (N = 97)

Records excluded after full-text review (N = 46)

- No full-text available (N = 10)
- Replicating results available in another paper (N = 1)
- No relevant health effects quantified (N = 20)
- Not an empirical paper (N = 7)
- Not a relevant intervention (N = 8)

Additional searches (N = 10)

Included

Sample of included studies (N = 61)

[a]These searches included manual searches of the MDRC, Abt Associates, Mathematica, and RTI websites and the Digest of Social Experiments published by the Urban Institute in 2004.
[b]Studies were excluded at this stage of the review for the following reasons: study was not a relevant intervention; no relevant health effects were quantified; study was not based in the United States; no abstract was electronically available (eg, conference proceedings or commentaries).

interquartile range [IQR] = 3,375). Two-arm social experiments were the most common design (90% of RCTs). The median duration of intervention was 48 months (IQR = 24 months) and the median follow-up period after randomization for health outcomes was 54 months

Table 1. Sample Study Characteristics (N = 61 Studies)

Characteristic	Value
Published in peer-reviewed journal, n (%)	40 (65)
Sample size, median (IQR)	1,866 (3,375)
Intervention duration in months, median (IQR)	48 (24)
Follow-up duration in months, median (IQR)	54 (84)
Multiple follow-ups, n (%)	17 (28)
Subgroups analyzed, n (%)	
By gender	6 (10)
By age	7 (12)
Other[a]	6 (10)
Two-arm RCT, n (%)	55 (90)
Three-arm RCT, n (%)	6 (10)
>1 health outcome reported, n (%)	45 (74)
No. of health outcomes reported, mean (SD)	6 (9)
Intervention domain	
Early life and education	19 (31)
Income supplementation and maintenance	9 (15)
Employment	14 (23)
Housing and neighbourhood	10 (16)
Health insurance	9 (15)

Abbreviations: IQR, interquartile range; RCT, randomized controlled trial; SD, standard deviation.
[a] "Other" category includes subgroup analyses by ethnicity, mother's age, and marital status, and at-risk populations (eg, people with criminal records or substance abuse histories).

(IQR = 84 months). Multiple follow-up measures were included in 28% of studies. Three-quarters of studies reported more than one health outcome, with a mean of 6 (SD = 9). Subgroup analyses were conducted in a minority (12%) of studies, most often by including multiple age groups or cohorts. Eligible social experiments covered five policy domains: early life and education (31% of studies), income supplementation and maintenance (15%), employment and welfare-to-work (23%), housing/neighborhood (16%), and health insurance (15%).

The Appendix Table details the 38 interventions included in the review in terms of objective, study design, setting, participants, and the effect of the policy on primary socioeconomic outcomes of interest. Eleven of the studies that we included tested interventions in the domain

of early life and education. These policies included intensive preschool and early education programs,[46–54] Head Start,[55,56] additional schooling support for people who dropped out of school and disadvantaged youths,[57–60] smaller class sizes,[61,62] alternative schools,[63] and vocational training.[64] All programs targeted at-risk children and youth. "At risk" was defined based on a risk measure, schooling level for age, health status, or a proxy measure for poverty. With the exception of the Quantum Opportunity Program Demonstration[58] and, to a certain extent, the Alternative Schools Demonstration Program,[63] all interventions reported positive impacts on a range of socioeconomic outcomes such as IQ scores, schooling duration, educational attainment, employment, and earnings.

Seven of the included studies examined the social and health impacts of income maintenance and supplementation programs. These included conditional cash transfers,[65,66] self-sufficiency programs,[67] and negative income taxes.[68–73] Low-income households were the target population of all these experiments. The Gary Experiment focused specifically on low-income African American families.[70] Most income experiments were associated with improvements in income and employment. Work Rewards in New York City was associated with moderate increases in employment and income only among participants who were unemployed at enrollment.[67] The findings of the negative income tax experiments remain heavily contested due to methodological concerns, but the interventions seem to be associated with a modest reduction in working hours and a decrease in marital stability.[74]

Thirteen of the included studies focused on employment and welfare-to-work: team-based supported employment,[75] job training programs,[76–78] employment support services,[79–84] and limits on welfare benefits coupled with income disregards and employment support services.[32,85–87] A third of the studies targeted respondents with mental health problems, while the other interventions focused more generally on people eligible for welfare benefits. About 75% of the tested welfare-to-work programs induced modest increases in earnings and reductions in welfare reliance.

Four interventions targeted housing/neighborhood changes: integrated clinical and housing services,[88,89] housing vouchers with[90–95] or without[96] a requirement to move to a higher-income neighborhood, and rental assistance.[97] Two of these studies intervened in specific population subgroups (homeless veterans with health conditions[88,89] and homeless people living with HIV/AIDS[97]). All trials were associated

with improved socioeconomics outcomes, measured as housing mobility, housing quality, or reduction in homelessness.
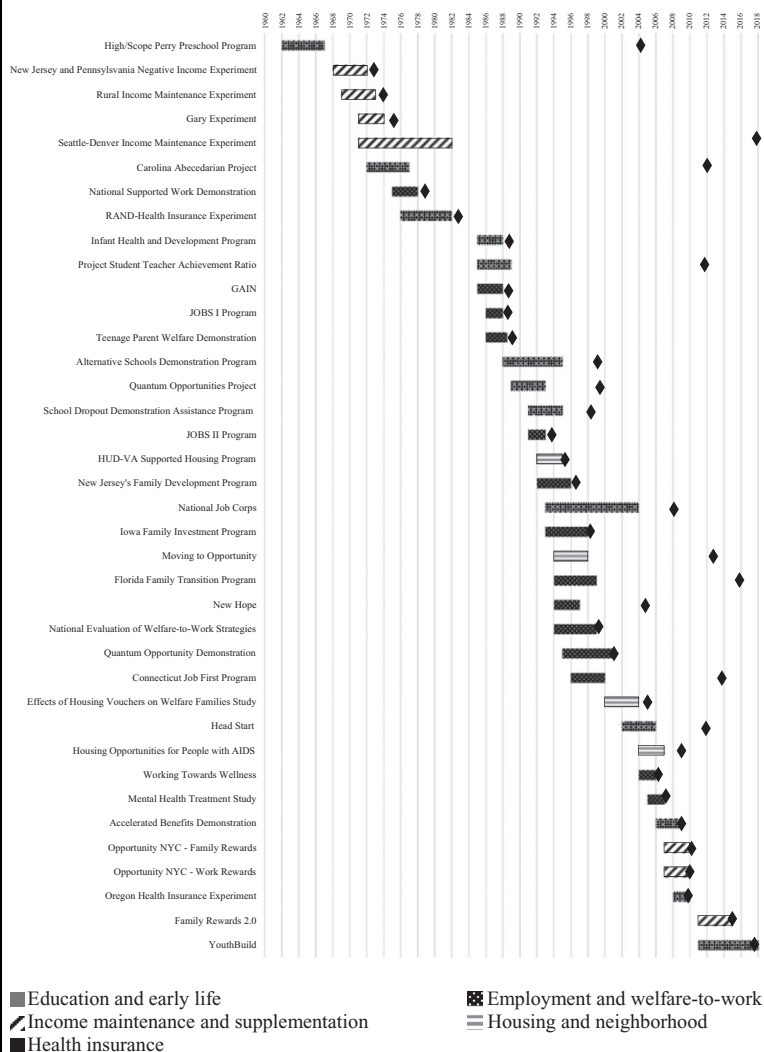
Finally, three trials evaluated the effect of providing health insurance coverage. The RAND-Health Insurance Experiment studied different tiers of private health insurance,[98–100] the Oregon Health Insurance Experiment investigated expansion of Medicaid coverage for low-income adults,[101–103] and the Accelerated Benefits Demonstration assessed the provision of health care benefits to new Social Security Disability Insurance (SSDI) beneficiaries.[105,106] The RAND[98–100] and the Oregon[101–103] health insurance experiments focused on uninsured low-income respondents, and both showed that medical care use increased when participants were able to pay for medical care, with the Oregon lottery nearly eliminating catastrophic out-of-pocket medical expenditures. The Accelerated Benefits study targeted new SSDI recipients without health insurance who were not yet eligible for Medicare, and the intervention was associated with an increase in job seeking.[105,106]

Figure 2 provides an overview of trial duration and health follow-up between 1960 and 2018. The first trial in our sample started in 1962; the latest was launched in 2007. There are three notable clusters of studies. The first corresponds to the four federally sponsored negative income tax experiments implemented between 1968 and 1982.[30] As indicated in the Appendix Table, these experiments varied in size. The second cluster includes randomized welfare-to-work experiments implemented between 1985 and 1988. Partly stimulated by the Omnibus Budget Reconciliation Act of 1981, these trials explored how to increase the earnings and employment rates of welfare recipients while decreasing their reliance on transfer payments. The third cluster corresponds to a range of federally funded welfare experiments that took place in the early to mid-1990s. These sought to study the effects of income disregards coupled with time limits on individual or family eligibility for welfare benefits.[36]
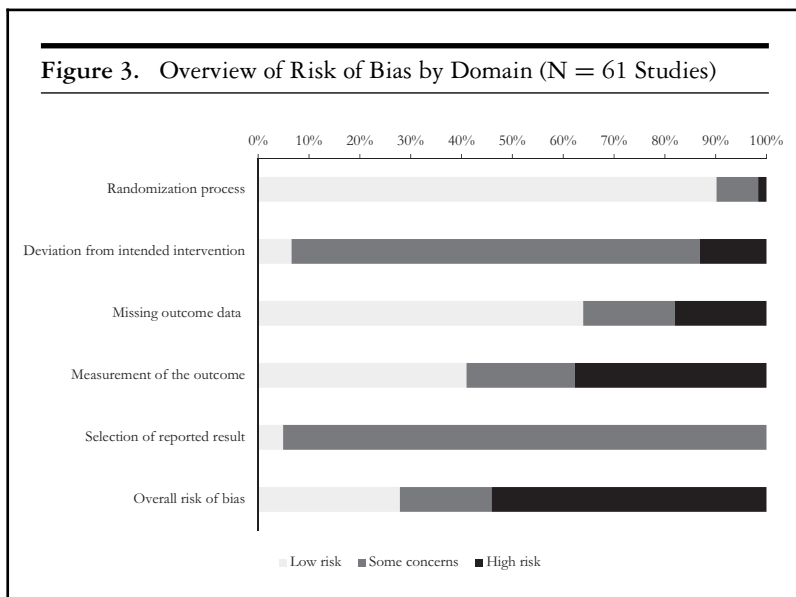
Risk of Bias

Figure 3 presents our domain-specific and overall risk of bias assessment. See Online Appendix 3 for risk of bias for each study. Overall, we judged 28% of the studies we reviewed to be at low risk of bias, 54% were at high risk of bias, and 18% had some concerns. These results were driven by higher risk of bias in the measurement of the outcome (38% of studies) and missing outcome data domains (18% of studies). By design, social

Figure 2. Trial Duration and Timing of Health Measurement (N = 38 Interventions)[a]

High/Scope Perry Preschool Program
New Jersey and Pennsylvania Negative Income Experiment
Rural Income Maintenance Experiment
Gary Experiment
Seattle-Denver Income Maintenance Experiment
Carolina Abecedarian Project
National Supported Work Demonstration
RAND-Health Insurance Experiment
Infant Health and Development Program
Project Student Teacher Achievement Ratio
GAIN
JOBS I Program
Teenage Parent Welfare Demonstration
Alternative Schools Demonstration Program
Quantum Opportunities Project
School Dropout Demonstration Assistance Program
JOBS II Program
HUD-VA Supported Housing Program
New Jersey's Family Development Program
National Job Corps
Iowa Family Investment Program
Moving to Opportunity
Florida Family Transition Program
New Hope
National Evaluation of Welfare-to-Work Strategies
Quantum Opportunity Demonstration
Connecticut Job First Program
Effects of Housing Vouchers on Welfare Families Study
Head Start
Housing Opportunities for People with AIDS
Working Towards Wellness
Mental Health Treatment Study
Accelerated Benefits Demonstration
Opportunity NYC - Family Rewards
Opportunity NYC - Work Rewards
Oregon Health Insurance Experiment
Family Rewards 2.0
YouthBuild

■ Education and early life   ▨ Employment and welfare-to-work
▨ Income maintenance and supplementation   ☰ Housing and neighborhood
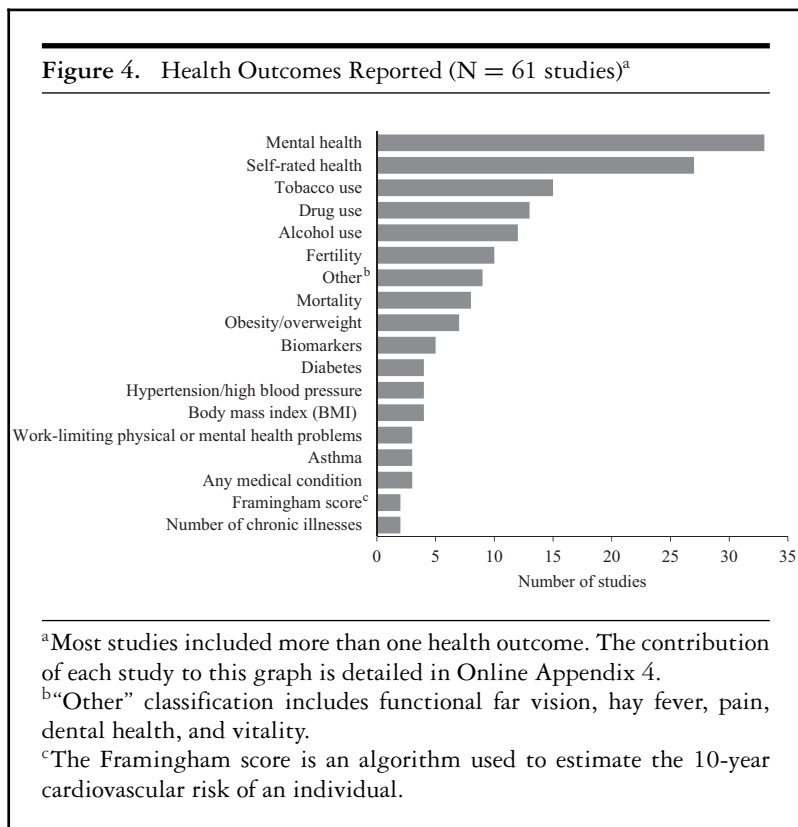■ Health insurance

[a]For each trial, the shaded area indicates the trial duration in years, and the diamond (♦) indicates the latest health outcome measurement available.

Figure 3. Overview of Risk of Bias by Domain (N = 61 Studies)

experiments are not blinded. Studies were categorized as at high risk of bias in the measurement of the outcome if they only included self-reported outcomes because assessment might be biased by participants' knowledge of their intervention group. It is also worth noting that only three studies[101–103] were registered or had a pre-analysis plan to which we could refer in our assessment.

Impacts on Health

The social experiments included in our review measured a wide variety of health outcomes (Figure 4). The most commonly reported outcome was mental health, which was measured in 33 studies. However, the evaluations of mental health used very different measures, such as signs of major depression, mental disorders, or established psychological distress scales (see Online Appendix 4 for an overview of all health measures per study). About 61% of all health measures were collected on survey instruments (either using validated scales or single-item self-report). Biomarkers and mortality (objective health outcomes) were reported in five and eight studies, respectively.

Figure 4. Health Outcomes Reported (N = 61 studies)[a]

[a]Most studies included more than one health outcome. The contribution of each study to this graph is detailed in Online Appendix 4.
[b]"Other" classification includes functional far vision, hay fever, pain, dental health, and vitality.
[c]The Framingham score is an algorithm used to estimate the 10-year cardiovascular risk of an individual.

We extracted 451 parameter estimates (the values assigned to outcome measures) from the 38 experiments in our sample. We were able to calculate post hoc statistical power for 75% of these estimates (N = 336). Most of the parameter estimates were underpowered to detect an effect of the intervention on health: only 23% had power of 80% or greater (see Online Appendix 5). Figure 5 summarizes the health effects of all interventions broken down by policy domains and across all extracted parameter estimates (Panel A) vs adequately powered estimates (Panel B). Across all parameter estimates, 71% showed a null effect on health, 26% showed a positive health effect, and 3% showed a worsening of heath in the treatment group relative to the control group. Online Appendix 4 details the contribution of each study to these findings.

Figure 5. Overview of the Effects of Social Experiments on Health Outcomes by Policy Domain[a]

Panel A. Effects across All Health Estimates (N = 451)
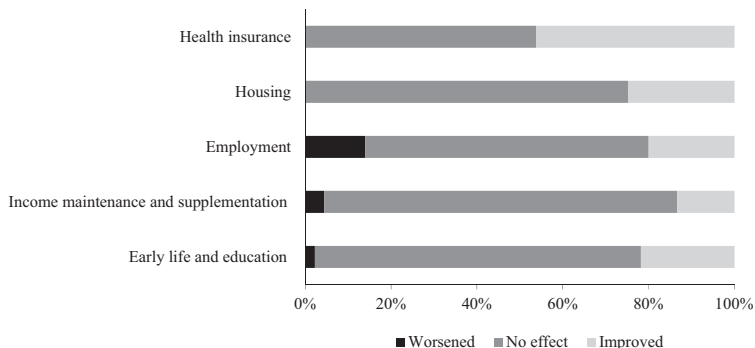
*Panel B. Effects across Adequately Powered Health Estimates (N = 86)*
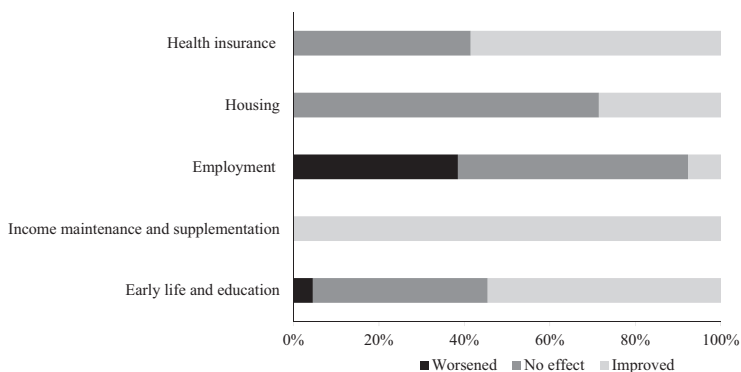
[a]No effect indicates a confidence interval that crosses 0 or $P \geq .05$. In Panel A, 451 estimates come from the 61 studies that compose our sample: 153 estimates in the early life and education domain, 45 in the income supplementation and maintenance domain, 52 in the employment and welfare-to-work domain, 110 in the housing and neighborhood domain, and 91 in health insurance domain. In Panel B, the 86 estimates are from those studies that are adequately powered (power $\geq 80\%$) to detect a health effect: 22 estimates in the early life and education domain, 3 in the income supplementation and maintenance domain, 13 in the employment and welfare-to-work domain, 7 in the housing and neighborhood domain, and 41 in the health insurance domain.

Among adequately powered studies (Figure 5 Panel B), 44% of the parameter estimates showed a null effect, 49% showed a positive health effect and 7% showed a negative effect compared to the control group. Negative health effects were mainly concentrated among time-limited welfare-to-work interventions (38% of all estimates for that policy domain). Most estimates in early life and education and health insurance interventions reported a positive health effect (55% and 59%, respectively). All estimates for income maintenance and supplementation experiments corresponded to an improvement in health, but we only had three adequately powered estimates in that domain. Of the adequately powered models in the housing category, 71% reported a null association with health. When we focused on adequately powered models among interventions that improved their primary socioeconomic outcomes (Online Appendix 6), we obtained similar results as in Figure 5 Panel B.

Figure 6 presents meta-analysis results for comparable subgroups of studies that measured good or excellent self-rated health, smoking status, and obesity. When pooling estimates from multiple RCTs, early life and education interventions were found to have a modest beneficial effect on smoking status (OR $= 0.92$, 95% CI 0.86-0.99, I^2 0.0%) but no effect on self-rated health (OR $= 1.02$, 95% CI 0.90-1.15, I^2 35.4%) or obesity (OR $= 1.10$, 95% CI 0.63-1.93, I^2 0.0%). Income maintenance and supplementation interventions were associated with an improvement in self-rated health (OR $= 1.20$, 95% CI 1.06-1.36, I^2 0.0%) but not with smoking status (OR $= 0.99$, 95% CI 0.71-1.37, I^2 85.2%) or obesity (OR $= 0.99$, 95% CI 0.90-1.09, I^2 0.0%). Pooled estimates for welfare-to-work interventions indicated that these experiments led to lower odds of reporting good or excellent self-rated health (OR $= 0.77$, 95% CI 0.66-0.90, I^2 78.4%). Housing/neighborhood experiments were not associated with self-rated health (OR $= 1.04$, 95% CI 0.95-1.14, I^2 0.0%) or smoking status (OR $= 1.06$, 95% CI 0.98-1.16, I^2 0.0%). Health insurance interventions improved self-rated health (OR $= 1.38$, 95% CI 1.10-1.73, I^2 73.3%) but had no impact on obesity (OR $= 1.03$, 95% CI 0.97-1.09, I^2 0.0%).

Table 2 summarizes the trial characteristics associated with reported findings across the 451 estimates. Publication in a peer-reviewed journal was associated with higher odds of statistical significance (OR $= 1.89$, 95% CI 1.04-3.45) and health improvement (OR $= 1.92$, 95% CI 1.02-3.61). Improvement in the primary socioeconomic outcome reported in the trial was associated with higher odds of health improvement

Figure 6. Meta-analyses of the Effects of Social Experiments on Health[a]
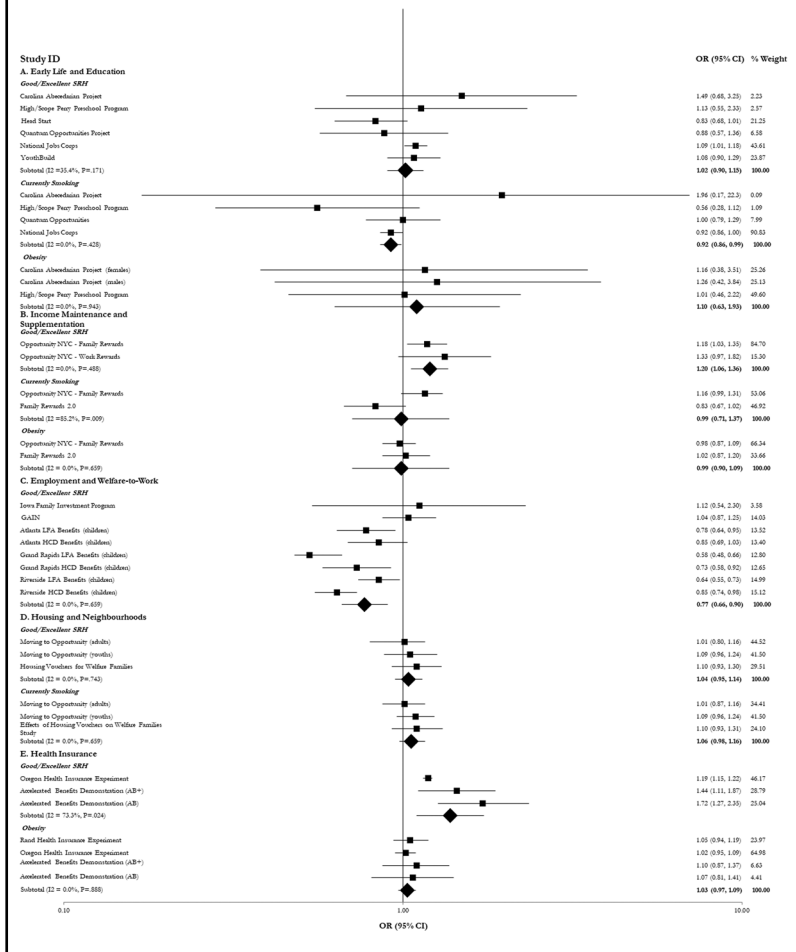
Abbreviations: AB, Accelerated Benefits; CI, confidence interval; HCD, Human Capital Development group (includes skill training and education); LFA, Labor Force Attachment group (focused primarily on job search); OR, odds ratio; SRH, self-rated health.
[a]Weights are from random effects analyses. The diamond shape corresponds to the pooled odds ratio.

Table 2. Association between Study Characteristics and Study Findings (N = 451 Estimates)[a]

	Odds Ratio (95% CI)	
	Statistically Significant Finding	Health Outcome Improved
Published in peer-reviewed journal	1.89[b] (1.04, 3.45)	1.92[b] (1.02, 3.61)
Year published	1.01 (0.97, 1.04)	0.99 (0.96, 1.02)
Sample size	1.00 (0.99, 1.00)	1.00 (0.99, 1.00)
Duration of intervention	0.99 (0.97, 1.01)	0.99 (0.97, 1.00)
Duration of follow-up	0.99 (0.99, 1.00)	0.99 (0.99, 1.00)
Primary socioeconomic outcome improved	1.25 (0.55, 2.88)	5.26[c] (1.44, 19.24)
Multiple study sites	0.85 (0.41, 1.76)	0.70 (0.33, 1.49)

[a] Robust standard errors are clustered at the study level (N = 61).
[b] $P < .05$.
[c] $P < .01$.

(OR = 5.26, 95% CI 1.44-19.24). Publication year, sample size, the duration of intervention and follow-up, and whether the trial was a multisite study were not associated with statistical significance or improvements in health measures.

Discussion

The "social policy is health policy" trope has led to the investment of tens of billions of health dollars in social policies such as education or housing.[13–15] We sought to review the experimental evidence linking social interventions to health outcomes and found that remarkably little has been done to understand whether social policies might actually

improve health. We identified 61 studies that detail the health effects of 38 randomized interventions in five policy domains and extracted 451 single estimates from these studies.

The social experiments included in our review spanned the period 1962-2018 and featured a range of policies, analytic approaches, target groups, and health outcomes. We provided both a narrative report of these findings and a meta-analysis. The narrative report is important because many studies and outcomes within these studies could not be included in the meta-analysis. The meta-analysis suggests what might be generally expected from similar investments within each category.

Our review allows us to draw several conclusions. First, over three-quarters of the models extracted were underpowered to detect a health effect. Statistical power is a critical parameter to assess the scientific and policy value of a study.[107] Low statistical power is associated with high rates of false negatives—the inability to detect an effect when it is there.[108] Because factors such as duration of follow-up can impact the chance of a false negative for health outcomes in social policy experiments, statistical power is only one clue used to determine the chance of a false null finding.

Second, among adequately powered studies, 49% of all measures showed improved health, 44% showed a null effect, and 7% showed an adverse effect on health (Figure 5 Panel B). When we broke down these numbers by policy domains, multiple interventions seemed to be promising ways to improve population health. Over half of the early life and education interventions had a positive effect on health. These studies showed broad improvements in a range of health outcomes.[47,49,51,52] Among those studies for which a meta-analysis could be performed, the early life and education programs were associated with an 8% reduction in smoking prevalence. Although this is a fairly small percentage, its implications may be substantial because smoking is a powerful determinant of population health. It is likely a strong proxy measure of other forms of risk-taking behaviors, such as condom or seatbelt use,[109] and it is a powerful predictor of poor health and premature mortality.[6] Still, in our meta-analysis, education and early life interventions did not improve self-rated health. The explanation for that finding might have to do with follow-up time. One of the older studies—the Carolina Abecedarian Project—stopped following up with participants when they reached their mid-30s. At that time, biomarker data (eg, blood pressure

and other risk factors for cardiovascular disease) suggested that those in the control group had worse health than those in the intervention arm.[49] Both the Carolina Abecedarian Project and the High/Scope Perry Preschool Program had positive health impacts before age 40, findings that suggest the potentially powerful effects of education on health.[51,52] Income maintenance and supplementation programs were associated with a 20% increase in the odds of reporting good or excellent self-rated health. Health insurance programs had the largest effects on self-rated health—a 38% improvement in the odds of reporting good or excellent self-rated health.

Third, while many beneficial health effects were noted in individual studies, our meta-analysis of the available outcome measures across studies did not show consistent benefits. Income support and health insurance programs showed the most promise in improving population health. Moreover, social policies that showed benefits in the domains of smoking (eg, education) and self-rated health (eg, income support and health insurance) did not show benefits in most of the other health domains that we studied. The lack of statistically significant findings may be due to insufficient follow-up time, which can lead to false negatives even among studies that show adequate statistical power using the tests we employed. Linking social policy experiments to electronic health and mortality data could enable researchers to examine outcomes in later life stages, when chronic diseases start to appear.

Fourth, a small number of studies (7% of adequately powered estimates) found harmful effects of social policies on health. These were mainly concentrated among time-limited welfare-to-work interventions.[78] These interventions targeted low-income families eligible for Aid for Families with Dependent Children (AFDC) and were associated with increases in employment and earnings and decreases in welfare receipts. Muennig and colleagues found that Temporary Assistance for Needy Families (TANF) was associated with higher mortality rates than AFDC.[32] On deeper examination of TANF's distributional effects, the authors found that some families were unable to work when welfare benefits ran out (possibly because of large family sizes or poor health). Those who were unable to work became ineligible for welfare benefits after five years, and subsequently became destitute. Muennig and coauthors speculated that this adverse change could explain the increase in mortality.[32] It could also explain the reductions in children's self-rated health in this policy domain: in pooled estimates for the domain, there

was a 23% reduction in good or excellent self-rated health. This finding reinforces the hypothesis that there is a trade-off between work incentives (which take parents out of the home) and the health of families. One education study found an increase in mortality associated with reducing class sizes (relative to normal size classes), and the authors speculated that this was because children in smaller classes were more social and extroverted and therefore took greater risks in their youth.[61] Longer-term follow-up of this cohort could ultimately find mortality benefits.

Fifth, when an experiment had a positive effect on its primary socioeconomic outcome, that was associated with greater odds of reporting a positive effect on health. This conclusion strongly supports the broad hypothesis that "social policy is health policy." It also highlights the importance of testing social policies using experimental methods—some ineffective social policies are still in place, and many that were effective have not been funded or scaled up.[31] Still, enthusiasm for social investments as health policy may be tempered by the finding that some of the adequately powered studies we examined showed no health impact, which may indicate that they were true negative findings. Health care providers and insurers should not randomly invest in social policies and expect health benefits—they must invest in those that have been proven to work without causing negative unintended consequences.

Sixth, significant findings and positive effects on health were strongly associated with publication in a peer-reviewed journal (Table 2). Publication bias—the tendency for authors to only submit statistically significant findings to journals and for journals to only publish statistically significant findings—is a well-documented and serious problem across disciplines.[110,111] However, in our study, many of the null findings were underpowered to detect a positive effect; therefore, it may have been appropriate to not publish them in peer-reviewed journals—it is certainly counterproductive to publish false-negative results. Still, some of the unpublished null or negative findings in our review were adequately powered, suggesting that publication bias was present. This tendency to publish only positive findings might explain why the extant quasi-experimental literature shows that social policies produce a greater health benefit than we found in our review. For example, Hamad and colleague's meta-analysis of quasi-experimental evaluations of law-mandated increases in schooling duration found that these policies were associated with decreased probability of obesity (effect size: -0.20; 95% CI -0.40 to -0.02).[112] In our review, which included different

interventions than those evaluated by Hamad and coauthors, experiments in the early life and education domain were not associated with lower obesity risk.[49,52]

Finally, our study highlights the difficulties of capturing health outcomes in social experiments. In the Oregon Medicaid experiment, participants were provided with income protection, disease screening, and proven treatments. The authors' qualitative study suggested that the subjective impact of Medicaid on the participants' physical health was large and positive.[113] However, no quantitative impact of exposure to Medicaid on physical health was statistically significant. The negative income tax experiments further highlight the challenges of measuring health outcomes. In the Gary Experiment, we observed positive program impacts that were intuitive. At the time of the experiment in 1971, hunger was a problem in the United States, particularly among African Americans. The experiment showed that impoverished African American women who received additional income were more likely to have normal birth weight babies than were control group women.[70] On the other hand, the New Jersey and Pennsylvania Negative Income Experiment produced no change in medical care utilization or (in the short-term) health.[68] Together, these experiments suggest that the measured health benefits of social policies can vary greatly depending on who is being studied, what the outcome measure is, and where the study is taking place.

Some conditions must be considered before evaluating the health impacts of a social policy. First, the social policy must above all else accomplish its goal of improving social well-being. A number of experiments included in this review failed to improve their primary socioeconomic outcome. We also found in multivariate analyses that improvements in primary socioeconomic outcomes were strongly associated with the odds of reporting an improvement in health. Second, the improvements in social circumstances produced by the intervention must occur early in life or be of sufficient intensity or duration to overcome the effect of adverse economic circumstances accumulated over the life course. Poverty is believed to take a toll starting in utero,[114] but experiments targeting income or employment tend to target adults, making it potentially more challenging to "move the needle" on health. Third, the duration of follow-up must be adequate to capture the health outcome of interest. Effective social policy interventions may reduce depression rates over a

relatively short period because mental health states can change rapidly. However, other health outcomes may not be evident for many years. For example, early life poverty can damage the physiological systems that regulate blood pressure, blood sugar, cholesterol, and body weight[115] and it can take time for improved economic well-being to reverse or slow the physiological changes that lead to high cholesterol. Effects on blood sugar or cholesterol might be possible to measure in the intermediate term. However, it takes time for elevated cholesterol to lead to a heart attack, so physical health changes may become apparent only years after exposure to an effective intervention. Although we did not distinguish a systematic pattern in our analyses, it is likely that intensive, long-term interventions with considerable time to follow-up will be more beneficial for health.[116]

Limitations

Our study was subject to a number of limitations. First, we had no way of conducting a full quantitative assessment of the false-negative error rate. While the post hoc power analysis is informative, traditional power analyses based on outcome prevalence or sample variance cannot account for intervention intensity and duration. Results from our multivariate analyses indicated that intervention and follow-up durations were not associated with the odds of reporting a significant or positive health effect. The mean age and age distribution of the cohort also presented a special challenge. If a cohort is very young, the participants will likely be healthy for a relatively long follow-up period. If the cohort is older, they may have been exposed to a lifetime of poverty or other health risks prior to the intervention. However, we incorporated these issues indirectly in our post hoc power calculations: if the intervention was low intensity, brief, or conducted among young people, the observed effect size input to the power calculation should be smaller and thus reflect lower power.

Second, the studies that we examined did not use a consistent set of health outcome measures. A meta-analysis can only be conducted if the outcome measures are the same in two or more studies. Those that were used frequently, such as self-rated health, may not be appropriate for social policy RCTs because self-rated health takes a long time to change.[117]

Third, 35% of the studies we included were not peer reviewed. The organizations conducting these experiments tend not to be staffed by academics, and the authors of white papers targeted at local, state, and

federal policymakers have little incentive to also publish in peer-reviewed journals. In addition, we found that 54% of all studies included in the review were at high risk of bias. It is, for example, difficult to draw reliable conclusions from the large negative income experiments from the 1970s due to differential attrition in the treated and control groups and limitations in the analyses conducted.[118]

Fourth, although our search was extensive, verified by experts post hoc, and updated in January 2019, we may have missed relevant experiments. For example, our search strategy included broad search terms such as "health" and "morbidity," and it is possible that we missed papers focused on specific outcomes such as diabetes. Nevertheless, our efforts at unearthing studies were helped by the organizations that conduct social research for the government and comprehensive searches of the grey literature and reference lists.

Fifth, we focused only on studies conducted in the United States. While cross-national studies were not included in our review and meta-analysis, a lot can be learned from countries such as the United Kingdom, which conducted a large welfare-to-work experiment in the 2000s.[119]

Sixth, there was conceptual and methodological heterogeneity in our meta-analyses. Pooled estimates from diverse studies should be interpreted cautiously. Although the I^2 in most analyses was acceptable, the interventions within each policy domain were diverse. In addition, studies varied considerably by follow-up duration. Given the limited number of studies in each policy subdomain, we could not take this variability into account in the meta-analyses. Also, the distinct populations in the various selected studies may have had differential (even opposite) responses to social interventions. A related issue is that the low number of studies in each meta-analysis precluded us from thoroughly assessing publication bias and implementing meta-regression analyses. Although the I^2 values in some cases indicated significant heterogeneity, the small number of studies meant that we could not perform subgroup analyses.

The Importance of Social Experiments for Population Health

Experimental evaluations of social policy remain controversial, in part due to the perception among researchers and policymakers that RCTs of social interventions are extremely costly and potentially unethical.[120,121] Nevertheless, in the United States, a long tradition

of social experimentation has guided policymaking.[30] When findings are null or negative, billions of dollars that would otherwise have been spent on ineffective or harmful programs can be put to productive use. Wasting potentially life-saving resources is also an important ethical consideration. When findings are positive, they can provide a political shield against partisan attacks.

More can be done. For example, it is important to not only design evaluations with health in mind but also modify and retest policies until we get them right. Family Rewards 2.0 and Work Rewards are the only two policies in our sample that were revised based on previous experimentation and retested.[66] While we are advocating for revision and retesting, we acknowledge that fine-tuning existing policies is a big challenge; one reason that the popularity of social experiments declined after the 1990s may have been policy advocates' frustration with null and sometimes damaging effects reported in such studies. Adequately testing interventions in the field requires considerable investments of money and time, and these are long-term policy challenges that must be overcome if social experiments are to be considered a viable option.[30] It is particularly difficult for policymakers to acknowledge uncertainty about the effectiveness of social policies.[122] Petticrew and colleagues[122] point to the importance of fostering a culture of evidence production.

The evidence base in support of integrating the social determinants of health into health care delivery would be strengthened by employing best practices from across research disciplines.[124] Our findings indicate that this body of literature is vulnerable to publication bias. To address this shortcoming, investigators reporting on social experiments could adhere to the Consolidated Standards of Reporting Trials (CONSORT). These guidelines offer a standard way to report findings in a transparent and structured way, and they would considerably aid the critical appraisal and interpretation of the health effects of social experiments. If the studies we evaluated had followed such guidelines, we would, for example, have been better able to evaluate the risk of bias.[125] The preregistration of protocols required by the International Committee of Medical Journal Editors could also be extended to social experiments published in nonmedical journals. Only three of the studies that we examined had a prespecified and registered analysis plan.[101–103] We consequently could not fully evaluate the risk of bias due to selected reporting of outcomes in all studies. In addition, we were unable to examine the potential contribution of statistical analysis

misspecification to bias. The discussion of missing data was also often limited in the studies we analyzed. It is equally important for researchers to consider developing a list of comparable, validated, and objective health measures that can be tailored to the intervention type. Although the collection of biomarker data is expensive, they indicate subclinical biological processes that may be detected over a shorter term and they are not vulnerable to response bias. This is particularly important for social interventions, which by design cannot be double-blinded. The addition of health-related quality-of-life instruments would enable researchers to estimate the cost-effectiveness of nonmedical social policies. Recent research has shown, for example, that state supplements to the Earned Income Tax Credit are cost-effective.[26] These findings could be verified and expanded using experimental data. Where feasible, linking large-scale routinely collected data to past or ongoing experiments would support better-powered studies and enable researchers to garner valuable evidence on long-term health effects. Finally, making experimental data available when possible through one of the numerous data repositories available would improve the reproducibility of studies and confidence in their results.

Conclusion

After Virchow conducted a rigorous but associational study of the underlying causes of typhus, he concluded that the remedy should be to make heavy investments in social policies.[2] Since that early study, many thousands of associational studies have shown similar links between poverty and health. Based on this evidence, many governments, including the United States, have invested considerable resources in social programs using potential health care costs reductions as justification. Social experiments conducted in the United States since the 1960s constitute an untapped resource to inform policymaking. Interventions in the domains of early life and education, income, and health insurance are particularly promising as population health policies. However, this evidence is not without shortcomings: the vast majority of models were underpowered to detect health effects and many experiments were at high risk of bias. Future social experiments should incorporate validated health measures, and cross-learning from different disciplines would considerably strengthen the validity and policy relevance of these findings.

References

1. Tountas Y. The historical origins of the basic concepts of health promotion and education: the role of ancient Greek philosophy and medicine. *Health Promot Int*. 2009;24(2):185-192.
2. Virchow R. Notes on the typhoid epidemic prevailing in Upper Silesia. *Arch Pathologische Anatomic Physiologic Klinische Medizin*. 1849;2:143-322.
3. Feinstein JS. The relationship between socioeconomic status and health: a review of the literature. *Milbank Q*. 1993;71(2):279-322.
4. Adler NE, Ostrove JM. Socioeconomic status and health: what we know and what we don't. *Ann N Y Acad Sci*. 2006;896(1):3-15.
5. Marmot M. Social determinants of health inequalities. *Lancet*. 2005;365(9464):1099-1104.
6. Muennig P, Fiscella K, Tancredi D, Franks P. The relative health burden of selected social and behavioral risk factors in the United States: implications for policy. *Am J Public Health*. 2010;100(9):1758-1764.
7. Woolf SH. Social policy as health policy. *JAMA*. 2009;301(11):1166-1169.
8. Adler NE, Newman K. Socioeconomic disparities in health: pathways and policies. *Health Aff*. 2002;21(2):60-76.
9. Low MD, Low BJ, Baumler ER, Huynh PT. Can education policy be health policy? Implications of research on the social determinants of health. *J Health Polit Policy Law*. 2005;30(6):1131-1162.
10. Deaton A. Policy implications of the gradient of health and wealth. *Health Aff*. 2002;21(2):13-20.
11. Berkman LF. Unintended consequences of social and economic policies for population health: towards a more intentional approach. *Eur J Public Health*. 2011;21(5):547-548.
12. Roberts H. What is Sure Start? *Arch Dis Child*. 2000;82(6):435-437.
13. Kassler WJ, Tomoyasu N, Conway PH. Beyond a traditional payer—CMS's role in improving population health. *N Engl J Med*. 2015;372(2):109-111.
14. Sparer MS, Brown L, Muennig PA. (Re) defining the health care delivery system: the role of social services. KPMG Government Institute. May 2016. https://assets.kpmg/content/dam/kpmg/pdf/2016/07/9-salud-re-defining-the-healthcare-delivery-system-the-role-of-social-services.pdf. Accessed November 21, 2019.

15. Kuehn BM. Hospitals turn to housing to help homeless. *JAMA*. 2019;321(9):822-824.

16. Muennig PA, Reynolds M, Fink DS, Zafari Z, Geronimus AT. America's' declining well-being, health, and life expectancy: not just a white problem. *Am J Public Health*. 2018;108(12):1626-1631.

17. Centers for Medicare and Medicaid Services. National health expenditure data. https://www.cms.gov/files/document/highlights.pdf. Accessed January 19, 2020.

18. Salm M. Does job loss cause ill health? *Health Econ*. 2009;18(9):1075-1089.

19. Fuchs V. Reflections on the socio-economic correlates of health. *J Health Econ*. 2004;23:653-661.

20. Courtin E, Nafilyan V, Avendano M, et al. Longer schooling but not better off? A quasi-experimental study of the effect of compulsory schooling on biomarkers in France. *Soc Sci Med*. 2019;220(1):379-386.

21. Rehkopf DH, Strully KW, Dow WH. The short-term impacts of Earned Income Tax Credit disbursement on health. *Int J Epidemiol*. 2014;43(6):1884-1894.

22. Costello EJ, Erkanli A, Copeland W, Angold A. Association of family income supplements in adolescence with development of psychiatric and substance use disorders in adulthood among an American Indian population. *JAMA*. 2010;303(19):1954-1960.

23. Handley MA, Lyles CR, McCulloch C, Cattamanchi A. Selecting and improving quasi-experimental designs in effectiveness and implementation research. *Ann Rev Public Health*. 2018;39(1):5-25.

24. Preston SH, Elo IT. Anatomy of a municipal triumph: New York City's' upsurge in life expectancy. *Popul Dev Rev*. 2013;40(1):1-29.

25. Muennig P, Masters R, Vail D, Hakes J. The effects of New York City's' coordinated public health programmes on mortality through 2011. *Int J Epidemiol*. 2017;46(4):1239-1248.

26. Muennig PA, Mohit B, Wu J, Jia H, Rosen Z. Cost effectiveness of the Earned Income Tax Credit as a health policy investment. *Am J Prev Med*. 2016;51(6):874-881.

27. Easterbrook PJ, Gopalan R, Berlin J, Matthews DR. Publication bias in clinical research. *Lancet*. 1991;337(8746):867-872.

28. Greenburg D, Shroder M. *The Digest of Social Experiments*. 3rd ed. Washington, DC: Urban Institute Press; 2004.

29. Greenberg D, Linksz D, Mandell M. *Social Experimentation and Policy Making*. Washington, DC: Urban Institute Press; 2003.

30. Oakley A. Experimentation and social interventions: a forgotten but important history. *BMJ*. 1998;317:1239-1242.

31. Haskins R, Margolis G. *Show Me the Evidence: Obama's' Fight for Rigor and Results in Social Policy*. Washington, DC: Brookings Institution Press; 2014.

32. Muennig P, Rosen Z, Wilde ET. Welfare programs that target workforce participation may negatively affect mortality. *Health Aff (Millwood)*. 2013;32(6):1072-1077.

33. Kawachi I, Adler NE, Dow WH. Money, schooling, and health: mechanisms and causal evidence. *Ann NY Acad Sci*. 2010;1186:56-68.

34. Osypuk TL, Joshi P, Geronimo K, Acevedo-Garcia D. Do social and economic policies influence health? A review. *Curr Epidemiol Rep*. 2014;1(3):149-164.

35. Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ*. 2009;339:b2700.

36. Greenberg D, Shroder M. *Digest of Social Experiment*. Washington, DC: Urban Institute Press; 1997.

37. Truman BI, Smith-Akin CK, Hinman AR, et al. Developing the Guide to Community Preventive Services—overview and rationale. The Task Force on Community Preventive Services. *Am J Prev Med*. 2000;18(1 Suppl):18-26.

38. Olds DL, Kitzman H, Knudtson M, SAnson E, Smith J, Cole R. Impact of home visiting by nurses on maternal and child mortality: results of a two-decade follow-up of a randomized clinical trial. *JAMA Pediatr*. 2014;168(9):800-806.

39. Adler NE, Stewart J. Health disparities across the lifespan: meaning, methods, and mechanisms. *Ann NY Acad Sci*. 2010;1186(1):5-23.

40. Braveman P, Gottlieb L. The social determinants of health: it's time to consider the causes of the causes. *Public Health Rep*. 2014;129(1_suppl2):19-31.

41. Mazumder B, Miller S. The effects of the Massachusetts health reform on household financial distress. *Am Econ J: Econ Pol*. 2016;8(3):284-313.

42. Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ*. 2019;366:l4898.

43. Becker GS, Murphy KM, Tamura R. Human capital, fertlity and economic growth. *J Polit Econ*. 1990;98(5):12-37.

44. Higgins JPT, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *J R Stat Soc Ser A Stat Soc*. 2009;172(1):137-159.

45. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med*. 2002;21(11):1539-1558.

46. Campbell F, Ramey CT, Pungello E, Sparling J, Miller-Johnson S. Early childhood education: young adult outcomes from the Abecedarian project. *Appl Dev Sci*. 2002;6(1):42-57.

47. Campbell F, Wasik BH, Pungello E, et al. Young adult outcomes of the Abecedarian and CARE early childhood educational interventions. *Early Child Res Q*. 2008;23:452-466.

48. Campbell F, Pungello E, Burchinal M, et al. Adult outcomes as a function of an early childhood educational program: an Abecedarian Project follow-up. *Dev Psychol*. 2012;48:1033-1043.

49. Campbell F, Conti G, Heckman JJ, et al. Early childhood investments substantially boost adult health. *Science*. 2014;343(6178):1478-1485.

50. McLaughlin AE, Campbell F, Pungello E, Skinner M. Depressive symptoms in young adults: the influences of the early home environment and early educational child care. *Child Dev*. 2007;78(3):746-756.

51. Muennig P, Robertson D, Johnson G, Campbell F, Pungello EP, Neidell M. The effect of an early education program on adult health: the Carolina Abecedarian project randomized controlled trial. *Am J Public Health*. 2011;101(3):512-516.

52. Muennig P, Schweinhart L, Montie J, Neidell M. Effects of a prekindergarten educational intervention on adult health: 37-year follow-up results of randomized controlled trial. *Am J Public Health*. 2011;99(8):1431-1437.

53. Brooks-Gunn J, McCormick MC, Shapiro S, Benasich A, Black GW. The effects of early education intervention on maternal employment, public assistance and health insurance: the infant health and development program. *Am J Public Health*. 1994;84(6):924-931.

54. Martin A, Brooks-Gunn J, Klebanov P, Buka SL, McCormick MC. Long-term maternal effects of early childhood interventions: findings from the Infant Health and Development Program (IHDP). *J Appl Dev Psychol*. 2008;29(2):101-117.

55. Puma M, Bell S, Cook R, Heid C. *Head Start Impact Study: Final Report*. Washington, DC: US Department of Health and Human Services; 2010. https://www.acf.hhs.gov/sites/default/files/opre/hs_impact_study_final.pdf. Accessed November 20, 2019.

56. Puma M, Bell S, Cook R, et al. *Third Grade Follow-Up to the Head Start Impact Study*. Washington, DC: US Department of Health and Human Services; 2012. https://www.acf.hhs.gov/sites/default/files/opre/head_start_report.pdf. Accessed November 20, 2019.

57. Hahn A. *Evaluation of the Quantum Opportunities Program. Did the Program Work? A Report on the Post Secondary Outcomes and Cost-Effectiveness of the QOP Program (1989-1993)*. Waltham, MA: Brandeis University Heller Graduate School Center for Human Resources;1994. https://eric.ed.gov/?id=ED385621. Accessed November 21, 2019.

58. Schirm A, Stuart E, McKie A. *The Quantum Opportunity Program Demonstration: Final Impacts*. Washington, DC: Mathematica; 2006. https://www.mathematica.org/our-publications-and-findings/publications/the-quantum-opportunity-program-demonstration. Accessed November 21, 2019.

59. Schochet PZ, Burghardt J, Glazerman S. *National Job Corps Study: The Impacts of Job Corp on Participants' Employment and Related Outcomes*. Princeton, NJ: Mathematica; 2001. https://eric.ed.gov/?id=ED457358. Accessed November 21, 2019.

60. Dynarski M, Gleason P, Rangarajan A, Wood R. *Impacts of Dropout Prevention Programs*. Princeton, NJ: Mathematica; 1998. https://www.mathematica.org/our-publications-and-findings/projects/the-impacts-of-dropout-prevention-programs. Accessed November 21, 2019.

61. Muennig P, Johnson G, Wilde ET. The effect of small class sizes on mortality through age 29 years: evidence from a multicenter randomized controlled trial. *Am J Epidemiol*. 2011;173(12):1468-1474.

62. Wilde ET, Finn J, Johnson G, Muennig P. The effect of class size in Grades K-3 on adult earnings, employment and disability status: evidence from a multi-center randomized controlled trial. *J Health Care Poor Underserv*. 2011;22(4):1424-1435.

63. Dynarski M, Wood R. *Helping High-Risk Youth: Results from the Alternative Schools Demonstration Program*. Princeton, NJ: Mathematica; 1997.

64. Miller C, Millenky M, Schwartz L, Goble L, Stein J. *Building a Future. Interim Impact Findings from the YouthBuild Evaluation*. New York City, NY: MDRC; 2016. https://www.mdrc.org/sites/default/files/YouthBuild_Interim_Report_2016_508.pdf. Accessed November 20, 2019.

65. Courtin E, Muennig P, Verma N, et al. Conditional cash transfers and health of low-income families in the US: evaluating the Family Rewards experiment. *Health Aff*. 2018;37(3):438-446.

66. Miller C, Miller R, Verma N, et al. *Effects of a Modified Conditional Cash Transfer Program in Two American Cities*. New York, NY: MDRC; 2016. https://www.mdrc.org/publication/effects-modified-conditional-cash-transfer-program-two-american-cities. Accessed November 21, 2019.

67. Verma N, Yang E, Nunez S, Long D. *Learning from the Work Rewards Demonstration. Final Results from the Family Self-Sufficiency in New York City*. New York, NY: MDRC; 2017. https://www.mdrc.org/publication/learning-work-rewards-demonstration. Accessed November 21, 2019.

68. Elesh D, Lefcowitz MJ. The effects of the New Jersey-Pennsylvania Negative Income Tax Experiment on health and health care utilization. *J Health Soc Behav*. 1977;18:391-405.

69. Kerachsky S. Chapter 5: State of health and the utilization of medical care. In: *The Rural Income Maintenance Experiment*. Washington, DC: US Department of Health, Education, and Welfare; 1976.

70. Kehrer B, Wolin C. Impact of income maintenance on low birth weight: evidence from the Gary experiment. *J Hum Resour*. 1979;14:434-462.

71. Keeley MC. The effects of negative income tax programs on fertility. *J Hum Resour*. 1980;15(4):675-694.

72. Price D, Song J. The long-term effects of cash assistance. Princeton University Industrial Relations Section working paper 621. June 2018. https://dataspace.princeton.edu/jspui/handle/88435/dsp01ng451m210. Accessed November 21, 2019.

73. Thoits P, Hannan M. Income and psychological distress: the impact of an income maintenance experiment. *J Health Soc Behav*. 1979;20(6):120-138.

74. Widerquist K. A failure to communicate: what (if anything) can we learn from the negative income tax experiments? *J Behav Exp Econ*. 2005;34(1):49-81.

75. Drake RE, Frey W, Bond GR, et al. Assisting Social Security Disability Insurance beneficiaries with schizophrenia, bipolar disorder, or major depression in returning to work. *Am J Psychiatry*. 2013;170(12):1433-1441.

76. Manpower Demonstration Research Corporation. *Summary and Findings of the National Supported Work Demonstration*. New York, NY: MDRC; 1980. https://www.mdrc.org/publication/summary-and-findings-national-supported-work-demonstration. Accessed November 21, 2019.

77. Riccio J, Friedlander D, Freedman S. *GAIN: Benefits, Costs and Three-Year Impacts of a Welfare-to-Work Program*. New York City,

NY: MDRC; 1994. https://www.mdrc.org/publication/gain-benefits-costs-and-three-year-impacts-welfare-work-program. Accessed November 21, 2019.

78. Hamilton G, Freedman S, Gennetian L, et al. *National Evaluation of Welfare-to-Work Strategies. How Effective Are Different Welfare-to-Work Approaches? Five-Year Adult and Child Impacts for Eleven Programs*. New York, NY: MDRC; 2001. https://www.mdrc.org/publication/how-effective-are-different-welfare-work-approaches. Accessed November 21, 2019.

79. Price R, Van Ryn M, Vinokur AD. Impact of a preventive job search intervention on the likelihood of depression among the unemployed. *J Health Soc Behav*. 1992;33(2):158-167.

80. Vinokur AD, Schul Y, Vuori J, Price RH. Two years after a job loss: long-term impact of the JOBS program on reemployment and mental health. *J Occup Health Psychol*. 2000;5(1):32-47.

81. Vinokur AD, Price RH, Caplan RD, van Ryn M, Curran J. The Jobs I preventive intervention for unemployed individuals: short- and long-term effects on reemployment and mental health. In: Murphy LR, Hurrell JJ, Sauter SL, Keita GP, eds. *Job Stress Interventions*. Washington, DC: American Psychological Association; 1995:125-138.

82. Aber JL, Brooks-Gunn J, Maynard RA. Effects of welfare reform on teenage parents and their children. *Future Child*. 1995;5(2):53-71.

83. Miller C, Huston AC, Duncan GJ, McLoyd GJ, Weisner TS. *New Hope for the Working Poor. Effects after Eight Years for Families and Children*. New York, NY: MDRC; 2008. https://www.mdrc.org/publication/new-hope-working-poor. Accessed November 21, 2019.

84. Butler D, Alson J, Bloom D, et al. *What Strategies Work for the Hard-to-Employ? Final Results of the Hard-to-Employ Demonstration and Evaluation Project and Selected Sites from the Employment Retention and Advancement Project*. New York, NY: MDRC; 2012. https://www.mdrc.org/publication/what-strategies-work-hard-employ. November 21, 2019.

85. Wilde ET, Rosen Z, Couch K, Muennig P. Impact of welfare reform on mortality: an evaluation of the Connecticut Jobs First program, a randomized controlled trial. *Am J Public Health*. 2014;104(3):534-538.

86. Fraker TM, Ross CM, Stapulonis RA, et al. *The Evaluation of Welfare Reform in Iowa: Final Impact Report*. Washington, DC:

Mathematica; 2002. https://www.mathematica.org/our-public ations-and-findings/publications/the-evaluation-of-welfare-re form-in-iowa-final-impact-report. Accessed November 21, 2019.

87. Rossi PH. *New Jersey's Family Development Program: A Overview and Critique of the Rutgers' Evaluation*. 1998. Baltimore, MD: University of Maryland.

88. Cheng A-L, Lin H, Kasprow W, Rosenheck R. Impact of supported housing on clinical outcomes analysis of a randomized trial using multiple imputation technique. *J Nerv Ment Dis*. 2007;195(1):83-88.

89. Rosenheck R, Kasprow W, Frisman L, Liu-Mares W. Cost-effectiveness of supported housing for homeless persons with mental illness. *Arch Gen Psychiatry*. 2003;60:940-951.

90. Fortson JG, Sanbonmatsu L. Child health and neighborhood conditions: results from a randomized housing voucher experiment. *J Hum Resour*. 2010;45(4):840-864.

91. Leventhal T, Brooks-Gunn J. Moving to Opportunity: an experimental study of neighborhood effects on mental health. *Am J Public Health*. 2003;93(9):1576-1582.

92. Sanbonmatsu L, Ludwig J, Katz L, et al. *Moving to Opportunity for Fair Housing Demonstration Program. Final Impacts Evaluation*. Washington, DC: US Department of Housing and Urban Development; 2011.

93. Ludwig J, Sanbonmatsu L, Gennetian L, et al. Neighborhoods, obesity, and diabetes: a randomized social experiment. *N Engl J Med*. 2011;365(16):1509–1519.

94. Osypuk TL, Schmidt NM, Bates LM, Tchetgen-Tchetgen EJ, Earls FJ, Glymour MM. Gender and crime victimization modify neighborhood effects on adolescent mental health. *Pediatrics*. 2012;130(3):472-481.

95. Osypuk TL, Tchetgen-Tchetgen EJ, Acevedo-Garcia D, et al. Differential mental health effects of neighborhood relocation among youth in vulnerable families. *Arch Gen Psychiatry*. 2012;69(12):1284-1294.

96. Wood R, Turnham J, Mills G. Housing affordability and family well-being: results from the housing voucher evaluation. *Hous Policy Debate*. 2008;19(2):367-412.

97. Wolitski RJ, Kidder DP, Pals SL, et al. Randomized trial of the effects of housing assistance on the health and risk behaviors of homeless and unstably housed people living with HIV. *AIDS Behav*. 2010;14:493-503.

98. Brook RH, Ware J, Rogers WH, et al. Does free care improve adults' health? Results from a randomized controlled trial. *N Engl J Med*. 1983;309:1426-1434.

99. Keeler EB, Brook RH, Goldberg GA, Kamberg CJ, Newhouse JP. How free care reduced hypertension in the health insurance experiment. *JAMA*. 1985;254(14):1926-1931.

100. Burciaga Valdez R, Brook RH, Rogers WH, et al. Consequences of cost-sharing for children's health. *Pediatrics*. 1985;75(5):952-961.

101. Finkelstein A, Taubman S, Wright B, et al. The Oregon Health Insurance Experiment: evidence from the first year. *Q J Econ*. 2012;127(3):1057-1106.

102. Baicker K, Taubman S, Allen H, Bernstein M, Gruber J, Newhouse JP. The Oregon experiment—effects of Medicaid on clinical outcomes. *N Engl J Med*. 2013;368(18):1713-1722.

103. Baicker K, Allen H, Wright B, Taubman S, Finkelstein A. The effect of Medicaid on management of depression: evidence from the Oregon Health Insurance Experiment. *Milbank Q*. 2018;96(1):29-56.

104. Wright B, Conlin AK, Allen H, Tsui J, Carlson MJ, Li HF. What does Medicaid expansion mean for cancer screening and prevention? Results from a randomized trial on the impacts of acquiring Medicaid coverage. *Cancer*. 2015;122(5):791-797.

105. Michalopoulos C, Wittenburg D, Israel DAR, et al. *The Accelerated Benefits Demonstration and Evaluation Project. Impacts on Health and Employment at Twelve Months*. New York, NY: MDRC; 2011. https://www.mdrc.org/publication/accelerated-benefits-demonstration-and-evaluation-project. Accessed November 21, 2019.

106. Michalopoulos C, Wittenburg D, Israel DAR, Warren A. The effects of health care benefits on health care use and health: a randoized trial for disability insurance beneficiaries. *Med Care*. 2012;50(9):764-771.

107. Ioannidis JPA, Stanley TD, Doucouliagos H. The power of bias in economics research. *Econ J*. 2017;127(605):236-265.

108. Ioannidis JPA. Why most published research findings are false. *PLoS Med*. 2005;2(8):e124.

109. Hersch J, Viscusi WK. Smoking and other risky behaviors. *J Drug Issues*. 1998;28(3):645-661.

110. Ioannidis JP, Trikalinos TA. An exploratory test for an excess of significant findings. *Clin Trials*. 2007;4(3):245-253.

111. Card D, Krueger A. Time-series minimum-age studies. A meta-analysis. *Am Econ Rev*. 1995;85(2):283-243.

112. Hamad R, Elser H, Tran DC, Rehkopf DH, Goodman SN. How and why studies disagree about the effects of education on health: a systematic review and meta-analysis of studies of compulsory schooling laws. *Soc Sci Med*. 2018;212:168-178.

113. Allen H, Wright B, Baicker K. New Medicaid enrollees in Oregon report health care successes and challenges. *Health Aff*. 2014;33(2):292-299.

114. Gluckman PD, Hanson MA, Cooper C, Thornburg KL. Effect of in utero and early-life conditions on adult health and disease. *N Engl J Med*. 2008;359(1):61-73.

115. Council on Community Pediatrics. Poverty and child health in the United States. *Pediatrics*. 2016;137(4):e20160339.

116. Doyle O, Harmon CP, Heckman JJ, Tremblay RE. Investing in early human development: timing and economic efficiency. *Econ Hum Biol*. 2009;7(1):1-6.

117. Schnittker J, Bacak V. The increasing predictive validity of self-rated health. *PloS One*. 2014;9(1):e84933.

118. Connor J, Rodgers A, Priest P. Randomised studies of income supplementation: a lost opportunity to assess health outcomes. *J Epidemiol Community Health*. 1999;53:725-730.

119. Riccio J, Bewley H, Campbell-Barr V, et al. *Implementation and Second-Year Impacts of Lone Parents in the UK Employment Retention and Advancement (ERA) Demonstration*. London, UK: Department for Work and Pensions; 2008. https://www.mdrc.org/publication/implementation-and-second-year-impacts-lone-parents-uk-employment-retention-and. Accessed November 21, 2019.

120. Thomson H, Hoskins R, Petticrew M, et al. Evaluating the health effects of social interventions. *BMJ*. 2004;328:282-285.

121. Macintyre S, Petticrew M. Good intentions and received wisdom are not enough. *J Epidemiol Community Health*. 2000;54(11):802.

122. Petticrew M, McKee M, Lock K, Green J, Phillips G. In search of social equipoise. *BMJ*. 2013;347:f4016.

123. Gottlieb LM, Alderwick H. Integrating Social and medical care: could it worsen health and increase inequity? *Ann Fam Med*. 2019;17(1):77-81.

124. Davey C, Hargreaves JR, Hanson K, et al. Trials and tribulations: cross-learning from the practices of epidemiologists and economists in the evaluation of public health interventions. *Health Policy Plan*. 2018;33(5):702-706.

125. Boone P, Eble A, Elbourne D. *Risk and Evidence of Bias in Randomized Controlled Trials in Economics*. London: Centre for Economic Performance, London School of Economics; 2013.

Address correspondence to: Emilie Courtin, Harvard Center for Population and Development Studies, 9 Bow St., Cambridge, MA 02138 (email: ecourtin@hsph.harvard.edu).

Appendix Table. Social Experiments Meeting Inclusion Criteria (N = 38 Interventions)

Name	Objective	Study Design	Setting and Participants	Effect on Intermediate Socioeconomic Outcomes
Early life and education (n = 11)				
Alternative Schools Demonstration Program[63]	To assess the effect of alternative high schools on dropout rates	Two-arm RCT: (1) admission to an alternative high school sponsored by the demonstration; (2) control group (not admitted)	In 7 urban school districts (Los Angeles, CA; Stockton, CA; Denver, CO; Wichita, KS; Cincinnati, OH; Newark, NJ; Detroit, MI) N = 924	The effect of the program varied depending on the study sites; overall, the program yielded mixed findings, with positive effects on dropouts in certain sites but no effect on employment.

Continued

Continued

Name	Objective	Study Design	Setting and Participants	Effect on Intermediate Socioeconomic Outcomes
Carolina Abecedarian Project[46–51]	To assess whether intensive early childhood education could improve school readiness	Two-arm RCT: (1) intensive educational childcare program from infancy to kindergarten entry; (2) control group (no intervention)	Children at risk of developmental delays or academic failure based on households scores on High Risk Index in North Carolina N = 52 to 148 depending on the outcome and study	Children assigned to the program had large IQ score improvements by age 3 years, higher reading and math abilities by age 15 years, lower rates of teen pregnancy and depression, and greater likelihood of attending college at age 21 years.

Continued

Name	Objective	Study Design	Setting and Participants	Effect on Intermediate Socioeconomic Outcomes
Head Start[55,56]	To determine the impact of Head Start on children's school readiness and parental practices that support children's development	Two-arm RCT: (1) Head Start program services; (2) control group (not in Head Start; could enroll in other early childhood programs or services)	3- and 4-year-old eligible children in 383 randomly selected Head Start centers. N = 2,259	Children in the program group had a better preschool experience and improved school readiness on a number of dimensions. The program had minimal impact on cognitive and socio-emotional outcomes of participating children.

Continued

Continued

Name	Objective	Study Design	Setting and Participants	Effect on Intermediate Socioeconomic Outcomes
High/Scope Perry Preschool Program[52]	To evaluate whether an intensive preschool program improved long-term outcomes	Two-arm RCT: (1) intensive 2-year program of 2.5 hours of interactive academic instruction with 1.5-hour weekly home visits; (2) control group (no intervention)	Children were recruited from low-income, predominantly African American neighborhoods in Ypsilanti, MI

N = 123 | Children in the treatment group were more likely to complete more schooling, have a stable family environment, and have higher earnings in adulthood. |

Continued

Continued

Name	Objective	Study Design	Setting and Participants	Effect on Intermediate Socioeconomic Outcomes
Infant Health and Development Program[53,54]	To assess the efficacy of early education on a range of parental and child outcomes	Two-arm RCT: (1) home visits from birth of a low-birth weight, premature child to the child's third birthday, and center-based child development programming in the second and third years of life; (2) control group receiving pediatric follow-up only	Low-birth weight, premature infants (birthweight <2500g and a gestational age <37 weeks) born in eight participating cities (Little Rock, AR; Bronx, NY; Cambridge, MA; Miami, FL; Philadelphia, PA; Dallas, TX; Seattle, WA; New Haven, CT) N = 875	Mothers in the intervention group were employed more months and returned to work sooner after births of their children. Children in the intervention group reported higher scores in math and reading.

Name	Objective	Study Design	Setting and Participants	Effect on Intermediate Socioeconomic Outcomes
Quantum Opportunities Project[57]	To evaluate whether a youth development demonstration can provide educational benefits	Two-arm RCT: (1) ≤750 hours of education, development, and service per year from ninth grade through high school; (2) control group (no intervention)	In-school youth or youth who had dropped out in San Antonio, TX; Philadelphia, PA; Milwaukee, WI; Saginaw, MI; Oklahoma City, OK N = 250	Those enrolled in the program were more likely to graduate from high school, enroll in college, and receive awards, and less likely to have children and drop out of school.

Name	Objective	Study Design	Setting and Participants	Effect on Intermediate Socioeconomic Outcomes
Quantum Opportunity Demonstration[58]	To assess the effect of offering intensive and comprehensive services on the high school graduate and postsecondary enrollment of at-risk youth	Two-arm RCT: (1) Five years of after-school program providing case management and mentoring, supplemental education, developmental activities, community service activities, supportive services, and financial incentives; (2) control group (no intervention)	Youth with low grades entering high schools with high dropout rates in Cleveland, OH; Fort Worth, TX; Houston, TX; Memphis, TN; Washington DC, Philadelphia, PA; Yakima, WA N = 1,069	The program did not impact high school graduation and postsecondary enrolment rates.

Continued

Name	Objective	Study Design	Setting and Participants	Effect on Intermediate Socioeconomic Outcomes
National Job Corps[59]	To assess the effect of a youth training program on employment and related outcomes	Two-arm RCT: (1) enrollment in Job Corps, which provided extensive education, training and other services; (2) control group, which could join other programs available in their communities	Youths ages 16 to 24 years from disadvantaged households (defined as living in a household that receives welfare or has income below the poverty level) N = 11,313	Participants in the intervention group had higher education attainment, higher employment, and higher earnings overall.

Continued

Name	Objective	Study Design	Setting and Participants	Effect on Intermediate Socioeconomic Outcomes
Project Student Teacher Achievement Ratio[61,62]	To assess the long-term effect of receiving instruction in small classes, regular-size classes, or regular-size classes with a certified teacher's aide	Three-arm RCT: (1) small class (13–17 students); (2) regular-size class (22–25 students); (3) regular-size class (22–25 students) with a certified teacher's aide	Children based in 79 Tennessee schools selected based on their size and willingness to participate in the study N = 11,601	Children assigned to small classes achieved higher test scores, had higher high school graduation rates, and were more likely to take college entrance examinations.
School Dropout Demonstration Assistance Program[60]	To assess the effect of dropout-prevention programs across the United States	Two-arm RCT: (1) program funded by US Department of Education; (2) control group (not enrolled in the program)	Children in middle schools in Albuquerque, NM; Atlanta, GA; Flint, MI; Long Beach, CA; Newark, NJ; Rockford, IL; Sweetwater, IN. N = 334	Only children enrolled in high-intensity programs had higher grades and were less likely to drop out.

Continued

Name	Objective	Study Design	Setting and Participants	Effect on Intermediate Socioeconomic Outcomes
YouthBuild[64]	To assess the effect of an intervention providing construction-related or other vocational training, educational services, counseling, and leadership development opportunities on employment outcomes	Two-arm RCT: (1) YouthBuild program; (2) control group (not enrolled)	Low-income youths aged 16-24 years who did not complete high school in 250 participating organization nationwide N = 3,929	Participants in the program group had higher participation in education and training, small increases in wages and earnings, and higher civic engagement.

Continued

Name	Objective	Study Design	Setting and Participants	Effect on Intermediate Socioeconomic Outcomes
Income maintenance and supplementation (n = 7)				
Family Rewards 2.0[66]	To assess the effect of a modified version of the Family Rewards model supplemented by staff guidance	Two-arm RCT: (1) Family Rewards 2.0 conditional cash transfer, with conditions in the domains of health care, education, and employment; (2) control group (no intervention)	Low-income households in the Bronx, NY, and Memphis, TN N = 2,400	Family Rewards 2.0 increased income and reduced poverty but also led to a reduction in employment covered by the unemployment insurance system, driven by the Memphis site.

Continued

Continued

Name	Objective	Study Design	Setting and Participants	Effect on Intermediate Socioeconomic Outcomes
Gary Experiment[70]	To test the effects of a negative income tax, consisting of an income guarantee accompanied by a tax rate on other income	Five-arm RCT: (1) 4 combinations of guarantees and tax rates; (2) control group	African American families with at least one child under the age of 18 years living in Gary, IN N = 404	The program was associated with a reduction in employment rates of males and single females, and an increased likelihood that teenager continued school.
New Jersey and Pennsylvania Negative Income Experiment[68]	To test the effects of a negative income tax, consisting of an income guarantee accompanied by a tax rate on other income	Nine-arm RCT: (1) 8 combinations of guarantees and tax rates; (2) control group	Low-income households with 1 nondisabled member between the ages of 18 and 59 years and ≥1 other member living in Trenton, NJ; Jersey City, NJ; Paterson, NJ; or Scranton, PA N = 1,357	The program was associated with an increase in parental unemployment and increased teenage high school graduation rates.

Continued

Continued

Name	Objective	Study Design	Setting and Participants	Effect on Intermediate Socioeconomic Outcomes
Opportunity NYC–Family Rewards[65]	To test the effect of a conditional cash transfer program in the US	Two-arm RCT: (1) Family Rewards conditional cash transfer, with conditions in the domains of health care, education, and employment; (2) control group (no intervention)	Low-income households in New York, NY N = 4,749	Family Rewards led to a significant increase in household income and a reduction in poverty and material hardship for the duration of the program but was not associated with an increase in parental employment.

Continued

Continued

Name	Objective	Study Design	Setting and Participants	Effect on Intermediate Socioeconomic Outcomes
Opportunity NYC–Work Rewards[67]	To test two strategies to increase employment and earnings of families receiving housing vouchers: the Family Self-Sufficiency program (FSS), which offers case management, and an enhanced version of this program (FSS + incentives), which includes cash incentives to encourage sustained full-time employment	Two-arm RCT: (1) FSS alone; (2) FSS plus special work incentives	Housing voucher recipients based in New York, NY N = 1,603	FSS and FSS + incentives did not increase employment or earnings overall and did not reduce receipt of housing assistance, poverty, or incidence of material hardship.

Continued

Name	Objective	Study Design	Setting and Participants	Effect on Intermediate Socioeconomic Outcomes
Rural Income Maintenance Experiment[69]	To test the effects of a negative income tax, consisting of an income guarantee accompanied by a tax rate on other income	Six-arm RCT: (1) 5 combinations of guarantees and tax rates; (2) control group	Rural low-income households, with a nondisabled man between the ages of 18 and 59 years living in Duplin (NC), Pocahontas (WV) or Calhoun (IA) counties N = 554	The program was associated with reduced family income and lower employment rates.

Continued

Name	Objective	Study Design	Setting and Participants	Effect on Intermediate Socioeconomic Outcomes
Seattle–Denver Income Maintenance Experiment[71–73]	To test the effect of both a negative income tax and subsidized vocational counseling and training	Three-arm RCT: (1) financial treatment at different levels of income guarantee and tax rate; (2) counseling/training treatment; (3) control group	Families meeting the following criteria: (1) either married or single with ≥1 dependent child under 18 years of age; (2) earning either <$9,000 or <$11,000 per year in 1971 dollars; (3) included a nondisabled head of household, aged 18 to 58 years N = 2,280 to 7,500 depending on the study	The programs were associated with longer unemployment spells, decreased earnings, and increased debt.

Continued

Name	Objective	Study Design	Setting and Participants	Effect on Intermediate Socioeconomic Outcomes
Employment and welfare-to-work (n = 13)				
Connecticut Jobs First Program[85]	To evaluate the effects of a 21-month limit on welfare benefits coupled with an employment mandate and incentives for finding and keeping a job	Two-arm RCT: (1) benefit cap of 21 months of welfare coverage, employment mandate, childcare assistance, earned income disregard, and 2 years of Medicaid eligibility after leaving welfare; (2) control group receiving AFDC benefits	Low-income families receiving AFDC in Connecticut N = 4,803	Participation in the program was associated with significant employment and income benefits.

Continued

Name	Objective	Study Design	Setting and Participants	Effect on Intermediate Socioeconomic Outcomes
Florida Family Transition Program[32]	To evaluate the effects of a 24- or 36-month limit on welfare benefits coupled with enhanced employment-related services such as education, job training, and job placement	Two-arm RCT: (1) time limits on welfare benefits and extra job training and case management; (2) control group receiving regular benefits and no additional job counseling	New AFDC applicants and existing AFDC recipients in Escambia and Alachua Counties, FL, ages 18-60 years, not in school, not working >30 hours per week, not disabled or caretakers for a disabled person, and not caring for a child age <6 months N = 3,324	Participants in the intervention group were more likely to find work, but their earnings did not tend to be greater than the total income from all sources received by the control group. Half of the intervention group remained unemployed during the program and reported relying on support from friends, family, and other programs such as food stamps.

Name	Objective	Study Design	Setting and Participants	Effect on Intermediate Socioeconomic Outcomes
GAIN[77]	To assess the impact of a welfare-to-work program aiming to reduce dependence and increase self-sufficiency by providing comprehensive support services	Two-arm RCT: (1) basic education and training; also, job search assistance for those who did not have a high school diploma or a GED, or job search assistance for those who were not screened as needing more basic education; (2) control group with access to other community services	Low-income families eligible to AFDC in 6 California counties N = 3,314	Participation in the program increased parental earnings and reduced welfare payments.

Name	Objective	Study Design	Setting and Participants	Effect on Intermediate Socioeconomic Outcomes
Iowa Family Investment Program[86]	To assess the impact of Iowa's new welfare-to-work program	Two-arm RCT: (1) comprehensive package of incentives to encourage self-sufficiency, including an earning disregard, participation in employment and training programs, and no restrictions on eligibility for 2-parent families; (2) control group receiving AFDC	Low-income families eligible for welfare benefits in Iowa N = 1,866	Participation in the program was associated with increases in employment, earnings, and incidence of domestic abuse.

Continued

Name	Objective	Study Design	Setting and Participants	Effect on Intermediate Socioeconomic Outcomes
JOBS I Program[79,80]	To evaluate the impact of an employment intervention on persons with different levels of risk factors for depression	Two-arm RCT: (1) eight 3-hour group job training sessions over a 2-week period; (2) control group (no intervention)	Unemployed people recruited from Michigan Employment Security Commission offices N = 928 to 1,122 depending on the study	The intervention primarily increased employment among participants at high-risk of depression.
JOBS II Program[81]	To evaluate the impact of an employment intervention on persons with different levels of risk factors for depression, with the addition of a screening tool to identify those at high-risk of losing jobs	Two-arm RCT: (1) five 3-hour group job training sessions over a 1-week period; (2) control group (no intervention)	Unemployed people recruited from Michigan Employment Security Commission offices N = 1,801	The intervention primarily increased employment among participants at high-risk of depression.

Continued

Continued

Name	Objective	Study Design	Setting and Participants	Effect on Intermediate Socioeconomic Outcomes
Mental Health Treatment Study[75]	To test whether supported employment and mental health treatments improve vocational and mental health recovery for people with psychiatric impairments	Two-arm RCT: (1) multifaceted intervention combining team-supported employment, systematic medication management, and other behavioral services; (2) control group receiving usual services	SSDI beneficiary with a primary diagnosis of schizophrenia or mood disorder, interested in gaining employment, age 18–55 years, and residing within 30 miles of one of the 23 study sites N = 2,238	The intervention was effective in assisting return to work and improving mental health and quality of life.

Continued

Name	Objective	Study Design	Setting and Participants	Effect on Intermediate Socioeconomic Outcomes
National Evaluation of Welfare-to-Work Strategies[77]	To assess the impact of labor-force attachment vs human capital development welfare-to-work programs on welfare recipients and their children	Two-arm RCT: (1) labor-force attachment or human capital development program; (2) control group (no intervention)	Low-income families eligible for AFDC in Atlanta, GA; Grand Rapids, MI; Riverside, CA; and Portland, OR. N = 2,938	Participation in employment-focused programs was associated with increases in employment and earnings and decreases in welfare receipt.

Continued

Continued

Name	Objective	Study Design	Setting and Participants	Effect on Intermediate Socioeconomic Outcomes
National Supported Work Demonstration[76]	To test the effect of employment and training programs on assisting hard-to-employ persons	Two-arm RCT: (1) work experience under conditions of gradually increasing demands, close supervision, and work in association with a crew of peers; (2) control group (no intervention)	Individuals with severe obstacles to obtaining employment and little recent work experience—ie, women receiving AFDC for many years, people with substance abuse histories, people with criminal records, and young adults who had dropped out of school—in 15 sites across the United States N = 6,616	The program was most effective in preparing women receiving AFDC and people with substance abuse histories for employment; it had no effects on those with criminal histories or the youth group.

Continued

Continued

Name	Objective	Study Design	Setting and Participants	Effect on Intermediate Socioeconomic Outcomes
New Hope[83]	To assess the impact of providing full-time workers with benefits and those unable to find full-time employment with support to find a job and referral to wage-paying community service job	Two-arm RCT: (1) earning supplement, health insurance, childcare assistance, and job access under a 30-hour work requirement; (2) control group (no intervention)	Low-income individuals age ≥18 years, willing and able to work full time, and living in a targeted Milwaukee, WI, neighbourhood N = 691	Participants were more likely to work, had higher incomes, and made more use of center-based childcare. Children from participating families had higher academic performance and test scores.

Name	Objective	Study Design	Setting and Participants	Effect on Intermediate Socioeconomic Outcomes
New Jersey's Family Development Program[87]	To assess the impact of New Jersey's new welfare-to-work program	Two-arm RCT: (1) family cap, earned-income disregard, no marriage penalty, increased benefits for two-parent households, extended Medicaid eligibility, employment support, and sanctions for noncompliance; (2) control group receiving AFDC benefits	Low-income families eligible for AFDC in New Jersey N = 1,232	The program reduced welfare dependency among participants, but its effects on employment and earnings are unclear.

Continued

Name	Objective	Study Design	Setting and Participants	Effect on Intermediate Socioeconomic Outcomes
Teenage Parent Welfare Demonstration[82]	To test the impact of a welfare-to-work program for teenage parents coupled with mandatory school and work requirements, and support services	Two-arm RCT: (1) mandatory school and work requirements enforced by financial sanctions; support services such as case management, parenting workshops, childcare assistance, and education and training opportunities; (2) control group receiving regular services	Teenage parents eligible for welfare in Chicago, IL, or New Jersey N = 3,867	The program increased attendance at school and job training programs and modestly increased employment rates.

Continued

Continued

Name	Objective	Study Design	Setting and Participants	Effect on Intermediate Socioeconomic Outcomes
Working Towards Wellness[84]	To evaluate the impact on employment and earnings of a care management program providing support for parents with depression who face barriers to employment	Two-arm RCT: (1) telephonic care management program, including education about depression, treatment, and monitoring of treatment adherence; (2) control group (no intervention)	Parents receiving Medicaid with symptoms of depression in Rhode Island N = 1,866	The program had no effect on employment.

Continued

Name	Objective	Study Design	Setting and Participants	Effect on Intermediate Socioeconomic Outcomes
Housing and neighbourhood (n = 4)				
Effects of Housing Vouchers on Welfare Families Study[96]	To measure the effects of vouchers on the well-being of low-income families eligible for or receiving TANF	Two-arm RCT: (1) experimental group receiving housing vouchers; (2) control group not receiving the housing voucher but remaining on the public housing authority's waiting list for regular assistance	Families eligible for Housing Choice Vouchers and current/former TANF recipients in Atlanta, GA; Augusta, GA; Fresno, CA; Houston, TX; Los Angeles, CA; and Spokane, WA N = 2,463	Vouchers reduced homelessness, crowding, household size, and incidence of living with family and friends. They also increased housing mobility and resulted in a small improvement in housing quality.

Continued

Name	Objective	Study Design	Setting and Participants	Effect on Intermediate Socioeconomic Outcomes
Housing Opportunities for People with AIDS[97]	To assess the effects of rental assistance on the health and risk behavior of homeless and unstably housed people living with HIV/AIDS	Two-arm RCT: (1) experimental group receiving immediate rental assistance with case management; (2) control group receiving customary housing services with case management	Adults age ≥18 years who were HIV-seropositive, homeless or at severe risk of homelessness, and with income <50% of median area income in Baltimore, MD; Chicago, IL; and Los Angeles, CA N = 630	Participation in the program was associated with improvements in housing status.

Continued

Continued

Name	Objective	Study Design	Setting and Participants	Effect on Intermediate Socioeconomic Outcomes
HUD-VA Supported Housing Program[88,89]	To evaluate the effects of providing integrated clinical and housing services to homeless veterans with psychiatric and/or substance abuse disorders	Three-arm RCT: (1) HUD-VA Supported Housing Program treatment group; (2) intensive case management without special access to Section 8 vouchers treatment group; (3) control group receiving standard VA homeless services	Veterans being treated at 19 participating VA medical centers located across the country. Veterans were eligible if they had been homeless for ≥ 1 month and had a diagnosis of a major psychiatric disorder and/or an alcohol or drug abuse disorder N = 272 to 370 depending on study	Participation in the program was associated with an increase in days housed and a reduction in homelessness.

Continued

Name	Objective	Study Design	Setting and Participants	Effect on Intermediate Socioeconomic Outcomes
Moving to Opportunity[90–95]	To evaluate the effects of providing housing vouchers to families living in public housing, with and without neighborhood restrictions	Three-arm RCT: (1) experimental group, receiving housing vouchers for use in low-poverty neighborhoods; (2) Section 8 treatment group receiving housing vouchers without neighborhood restrictions; (3) control group (no vouchers but eligible for public housing)	Children age ≤20 years from families living in high-poverty neighborhoods in Baltimore, MD; Boston, MA; Chicago, IL; Los Angeles, CA; and New York, NY N = 369 to 3,946 depending on study	Participation in the experimental group was associated with a lower likelihood of living in a high-poverty neighborhood, better housing quality, feeling safer in neighborhoods, and residing in less racially segregated neighborhoods.

Continued

Name	Objective	Study Design	Setting and Participants	Effect on Intermediate Socioeconomic Outcomes
Health insurance (n = 3)				
Accelerated Benefits Demonstration[105,106]	To test whether early access to health care and related services improves outcomes for new SSDI beneficiaries	Three-arm RCT: (1) AB group, which had access to health care benefits designed for the demonstration; (2) AB Plus group, which had access to the same health care benefits and voluntary services delivered by telephone to help navigate the health care system and return to work; (3) control group (no intervention)	New SSDI beneficiaries without health insurance who had ≥15 months until eligibility for Medicare; were aged 18–54 years; lived in one of the 53 metropolitan statistical areas with the highest number of new SSDI beneficiaries N = 1,226 or 1,531 depending on study	AB Plus services encouraged job seeking.

Continued

Name	Objective	Study Design	Setting and Participants	Effect on Intermediate Socioeconomic Outcomes
Oregon Health Insurance Experiment[101–104]	To assess the effect of expanding Medicaid coverage of low-income adults	Two-arm RCT: (1) Medicaid coverage; (2) control group (no intervention)	Low-income Oregon residents, aged 19-64 years, ineligible for other public health insurance and uninsured for the past six months. N = 4,166 to 74,922 depending on study	Medicaid coverage nearly eliminated catastrophic out-of-pocket medical expenditures.

Continued

Continued

Name	Objective	Study Design	Setting and Participants	Effect on Intermediate Socioeconomic Outcomes
RAND-Health Insurance Experiment[98–100]	To assess whether free medical care leads to better health than cost-sharing insurance plans	Two-arm RCT: (1) free medical care; (2) control group required to pay a share of their medical bills	Low-income individuals aged 14–61 years living in Seattle, WA; Dayton, OH; Fitchburg or Franklin County, MA; and Charleston or Georgetown County, SC. N = 1,844 to 6,622 depending on study	Cost sharing was associated with a reduction in medical care, in particular among poorest respondents.

Abbreviations: AFDC, Aid to Families with Dependent Children; FSS, Family Self-Sufficiency; GED, General Education Development; HUD, US Department of Housing and Urban Development; RCT, randomized controlled trial; SSDI, Social Security Disability Insurance; TANF, Temporary Assistance for Needy Families; VA, US Department of Veterans Affairs.

Supplementary Material

Additional supporting information may be found in the online version of this article at http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1468-0009:

Online Appendix 1. Eligibility Criteria
Online Appendix 2. Search Strategy
Online Appendix 3. Risk of Bias Assessment by Domain Across Individual Studies
Online Appendix 4. Overview of Health Outcomes and Intervention Effects Recorded in Each Study
Online Appendix 5. Post Hoc Power Calculations
Online Appendix 6. Health Effects Among Adequately Powered Studies That Also Had a Positive Effect on the Primary Socioeconomic Outcome of Interest