A Bayesian Reanalysis of Results from the Enhanced Services for the Hard-to-Employ Demonstration and Evaluation Project

OPRE Report 2012-36

June 2012

Office of Planning, Research and Evaluation (OPRE)
Administration for Children and Families
U.S. Department of Health and Human Services

Office of the Assistant Secretary for Planning and Evaluation
U.S. Department of Health and Human Services

A Bayesian Reanalysis of Results from the Hard-to-Employ Demonstration and Evaluation Project

OPRE Report 2012-36

June 2012

Author: Charles Michalopoulos, MDRC

Submitted to: Girley Wright, Project Officer
Office of Planning, Research and Evaluation
Administration for Children and Families
U.S. Department of Health and Human Services

Kristen Joyce and Laura Radel, Project Officers
Assistant Secretary for Planning and Evaluation
U.S. Department of Health and Human Services

Project Director: David Butler
MDRC
16 East 34 Street
New York, NY 10016

Contract Number: HHS-233-01-0012

Suggested citation: Charles Michalopoulos (2012). *A Bayesian Reanalysis of Results from the Hard-to-Employ Demonstration and Evaluation Project.* OPRE Report 2012-36. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Disclaimer: The views expressed in this publication do not necessarily reflect the views or policies of the Office of Planning, Research and Evaluation, the Administration for Children and Families, or the U.S. Department of Health and Human Services.

This report and other reports sponsored by the Office of Planning, Research and Evaluation are available at http://acf.gov.programs/opre/index.html.

MDRC is conducting the Enhanced Services for the Hard-to-Employ Demonstration and Evaluation Project under a contract with the Administration for Children and Families (ACF) and the Office of the Assistant Secretary for Planning and Evaluation (ASPE) in the U.S. Department of Health and Human Services (HHS), funded by HHS under a competitive award, Contract No. HHS-233-01-0012. The project officers are Girley Wright (ACF) and Kristen Joyce and Laura Radel (ASPE). Additional funding has been provided by the U.S. Department of Labor (DOL). DataStat, a subcontractor, fielded the six-month client survey for the Rhode Island site.

The findings and conclusions in this report do not necessarily represent the official positions or policies of HHS.

Dissemination of MDRC publications is supported by the following funders that help finance MDRC's public policy outreach and expanding efforts to communicate the results and implications of our work to policymakers, practitioners, and others: The Annie E. Casey Foundation, The George Gund Foundation, Sandler Foundation, and The Starr Foundation.

In addition, earnings from the MDRC Endowment help sustain our dissemination efforts. Contributors to the MDRC Endowment include Alcoa Foundation, The Ambrose Monell Foundation, Anheuser-Busch Foundation, Bristol-Myers Squibb Foundation, Charles Stewart Mott Foundation, Ford Foundation, The George Gund Foundation, The Grable Foundation, The Lizabeth and Frank Newman Charitable Foundation, The New York Times Company Foundation, Jan Nicholson, Paul H. O'Neill Charitable Foundation, John S. Reed, Sandler Foundation, and The Stupski Family Fund, as well as other individual contributors.

For information about MDRC and copies of our publications, see our Web site: www.mdrc.org.

This page intentionally left blank.

Abstract

Social policy evaluations usually use classical statistical methods, which may, for example, compare outcomes for program and comparison groups and determine whether the estimated differences (or impacts) are statistically significant — meaning they are unlikely to have been generated by a program with no effect. This approach has two important shortcomings. First, it is geared toward testing hypotheses regarding specific possible program effects — most commonly, whether a program has zero effect. It is difficult with this framework to test a hypothesis that, say, the program's estimated impact is larger than 10 (whether 10 percentage points, $10, or some other measure). Second, readers often view results through the lens of their own expectations. A program developer may interpret results positively even if they are not statistically significant — that is, they do not confirm the program's effectiveness — while a skeptic might interpret with caution statistically significant impact estimates that do not follow theoretical expectations.

This paper uses Bayesian methods — an alternative to classical statistics — to reanalyze results from three studies in the Enhanced Services for the Hard-to-Employ (HtE) Demonstration and Evaluation Project, which is testing interventions to increase employment and reduce welfare dependency for low-income adults with serious barriers to employment. In interpreting new data from a social policy evaluation, a Bayesian analysis formally incorporates prior beliefs, or expectations (known as "priors"), about the social policy into the statistical analysis and characterizes results in terms of the *distribution* of possible effects, instead of whether the effects are consistent with a true effect of zero.

The main question addressed in the paper is whether a Bayesian approach tends to confirm or contradict published results. Results of the Bayesian analysis generally confirm the published findings that impacts from the three HtE programs examined here tend to be small. This is in part because results for the three sites are broadly consistent with findings from similar studies, but in part because each of the sites included a relatively large sample. The Bayesian framework may be more informative when applied to smaller studies that might not be expected to provide statistically significant impact estimates on their own.

This page intentionally left blank.

Contents

This page intentionally left blank.

List of Exhibits

Table

Figure

This page intentionally left blank.

Introduction

Social policy evaluations almost always draw inferences using frequentist or classical statistical methods. An evaluator may, for example, compare outcomes for program and comparison groups and make statements about whether estimates are statistically significant, which means they are unlikely to have been generated by a program with a true effect of zero.

This approach has two important shortcomings. First, it is geared toward testing hypotheses regarding specific possible program effects. Most commonly, inferences are made about whether a program has zero effect. It is difficult with this framework to test a hypothesis that, say, the program's impact is larger than 10 (whether 10 percentage points, $10, or some other measure). Second, readers often view results through the lens of their own expectations. A program developer may interpret positive results that are not statistically significant as confirmation of the program's effectiveness, while a skeptic might interpret with caution statistically significant impact estimates that do not follow theoretical expectations.

An alternative paradigm that addresses both of these shortcomings is provided by Bayesian statistics. In interpreting new data from a social policy evaluation, a Bayesian analyst would formally incorporate prior beliefs (or "priors") about the effectiveness of the social policy into the statistical analysis — for example, a belief or expectation about how much a particular health care program could reduce the rate of hospital admissions. By formally specifying a range of prior beliefs, the analysis can show how robust the results are to different sets of beliefs rather than relying on readers to informally incorporate their own prior beliefs.

Bayesian and classical statistical inferences also differ. A typical program evaluation using the classical approach draws inferences concerning whether the program effect is likely to be zero. By contrast, the Bayesian approach characterizes results in terms of the *distribution* of possible effects — for example, greater than zero, or between five and ten, or in any other policy-relevant range.

Bayesian methods are not commonly used in drawing inferences from social policy experiments. A recent history of Bayesian thought describes two reasons for this.[1] First, except for the simplest applications, Bayesian analyses are computationally difficult and were infeasible until fast computers and simulation-based methods were developed. Second, it can be difficult to settle on a prior that is widely accepted. For example, the prior belief of a program developer may differ from the prior belief of a researcher. Even when priors are based on earlier research findings, there may be disagreements about which earlier studies should be used or how much weight different

[1] McGrayne (2011).

results should receive. Despite these hurdles, Bayesian statistical methods have been used more often in areas such as medical research, where several reviews of its use are available.[2]

This paper uses Bayesian methods to reanalyze results from the Enhanced Services for the Hard-to-Employ Demonstration and Evaluation Project, which tested interventions to increase employment and reduce welfare dependency for low-income adults with serious barriers to employment. The purpose of the paper is not to explore the utility of complicated Bayesian analyses that are now feasible, but to introduce some of the ideas underlying Bayesian analyses and to use relatively simple methods to aid in interpreting evidence from the evaluation's four program models.

The results reinforce the usefulness of two statistical methods that are already available within classical statistics. The first is meta-analysis, which looks across studies to provide stronger inferences than can be drawn from one study. The second is the use of confidence intervals to express the range of effects that are most consistent with new data, rather than characterizing results in terms of whether they are consistent with a true effect of zero.

This section introduces some concepts behind Bayesian analysis, while the remaining sections present the analysis for three programs in the Hard-to-Employ evaluation: the Rhode Island Working toward Wellness program (WtW), the New York Center for Employment Opportunities (CEO), and the Philadelphia Transitional Work Corporation (TWC).

An Introduction to Bayesian Ideas

Bayesian analysis has its foundation in Bayes' rule, which is illustrated here using an example from Gerd Gigerenzer's book *Calculated Risks*.[3] Consider the probability that a woman has breast cancer if she has a positive mammogram. According to Gigerenzer, 1 percent of women who are 40 years of age have breast cancer. A woman with breast cancer will screen positive on a mammogram 90 percent of the time, and overall, about 10 percent of women who have a mammogram screen positive.[4]

These numbers are used to create Table 1, which shows probabilities for the various combinations of whether a woman has a positive or negative mammogram (the rows) and

[2] Breslow (1990); Spiegelhalter, Freedman, and Parmar (1994); Spiegelhalter, Myles, Jones, and Abrams (2000).

[3] Gigerenzer (2002).

[4] Although Gigerenzer (2002) used real data for this example, current numbers may be different. The purpose of the example is to illustrate Bayesian ideas rather than to make a statement about the effectiveness of mammography.

Enhanced Services for the Hard-to-Employ Demonstration

Table 1

Probabilities of Women Age 40 Having Breast Cancer and Screening Positive on a Mammogram

Bayesian Reanalysis

Outcome (%)	Have Breast Cancer (%)	No Breast Cancer (%)	Total (%)
Positive mammogram	0.9	9.1	10.0
Negative mammogram	0.1	89.9	90.0
Total	1.0	99.0	100.0

SOURCE: Gigerenzer (2002).

whether she has breast cancer (the columns). The bottom row shows that, as noted above, 1 percent of 40-year-old women have breast cancer and 99 percent do not. The last column shows that 10 percent of women have positive mammograms while 90 percent do not. The four remaining cells show the full set of "joint probabilities" — that is, the probability that a woman has a positive mammogram *and* has breast cancer, the probability that a woman has a positive mammogram but does not have breast cancer, and so on.[5]

What are the chances that a woman has breast cancer if she screens positive? The table provides the answer: of every 1,000 women who have a mammogram, 100 (10 percent) will screen positive and 9 (0.9 percent) will screen positive and have breast cancer. Thus, the probability of having breast cancer given a positive mammogram is 9 in 100, or 9 percent; 91 percent of women with a positive mammogram do not have breast cancer.

Of course, factors such as family history and other risk factors would be taken into account in doing this kind of analysis. Nevertheless, this very simple example introduces some of the key ideas in Bayesian analysis. The probabilities in Table 1 represent the *prior* assessment of the various possible outcomes, determined before any new information is collected. The prior indicates that, in the absence of any other information, a 40-year-old woman has a 1 percent chance of having breast cancer. New data are collected, consisting of an observation of whether

[5] The numbers in Table 1 are derived from the probabilities presented in the previous paragraph. For example, since 90 percent of women with breast cancer will screen positive, that means 0.9 percent of all women will both screen positive and have breast cancer (that is, 90 percent of the 1 percent of women who have breast cancer).

a particular woman had a positive mammogram. The new data are used to update the prior into a "posterior" probability that the woman has breast cancer.

For discrete outcomes such as this, Bayes' rule can be written as:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

In this expression, *P(A|B)* is the probability that event *A* is true conditional on event *B* being true, *P(A)* is the probability that *A* is true, and *P(B|A)* and *P(B)* are defined analogously. In the breast cancer example, *A* is whether the woman has breast cancer and *B* is whether the woman has a positive mammogram. Thus, as shown in the equation above, the probability that the woman has breast cancer if her mammogram is positive *[P(A|B)]* is the probability her mammogram is positive given that she has breast cancer *[P(B|A)]* multiplied by the probability of having breast cancer *[P(A)]* and divided by the probability of testing positive *[P(B)]*. Using the numbers introduced at the beginning of this section, the probability a woman has breast cancer if she tests positive is 0.9 × 0.01/0.1 = 0.09, or 9 percent.

Extending Bayes' Rule to a Social Policy Evaluation

The Bayesian approach can also be applied to results from a social experiment. The steps are the same: a prior about the likely effects is developed, data are collected, and the two are combined to form a posterior distribution of the effect of the intervention.[6]

Forming a prior. The first step is to develop priors about the effectiveness of the program. Since the prior might affect the study's conclusions, Bayesian researchers have recommended using a range of priors to see how sensitive the results are to different priors.[7]

Several types of prior distributions have been discussed in the literature:

Uninformative prior. An uninformative prior can be used when there is little or no information on which to base expectations about the program's effectiveness. An uninformative prior might place an equal weight on a wide range of possible outcomes. For a study of any size, an uninformative prior yields conclusions that depend primarily on the new data that are collected.

[6] In the discrete expression of Bayes' rule above — *P(A|B) = [P(B|A) × P(A)]/P(B)* — *P(A)* is the prior probability, *P(B|A)* is the conditional distribution or likelihood of event *B* if *A* happens, and *P(A|B)* is the posterior probability of the event. For continuous outcomes such as the impact of an intervention, there is a continuous analog: the posterior probability distribution is proportional to the product of the prior probability distribution and the likelihood function related to the new data.

[7] Lilford and Braumholtz (1996).

Prior based on earlier studies. When similar studies have been conducted in the past, a prior might be based on results from those studies. A Bayesian analysis of an intervention to reduce spouse abuse in Colorado Springs, for example, developed priors using information from similar programs in Milwaukee and Omaha.[8] Although other studies are a natural source of priors, caution is in order if the interventions or studies differ. In that case, results from the other studies could be tempered with skepticism in forming the prior.[9] If the new intervention is a clear improvement over the old ones, optimism could be added to the previous results before forming a prior.

Prior based on expert opinions. When few earlier studies exist, a prior could be formed by asking a panel of experts about the likely effect of an intervention. In the Growth Restriction Intervention Trial, physicians involved in the study were asked to write their best guess of the study's effect on one key outcome and to use a histogram to indicate the distribution of their uncertainty.[10] In the Continuous Hyperfractionated Accelerated Radiotherapy Study in Head and Neck Cancer, 10 experts' opinions shaped the prior. Information from prior studies can also be supplemented with expert opinions, an approach that was taken in the Medical Research Council Neutron Therapy Trial.[11]

Skeptical and enthusiastic priors. Sensitivity tests can be done using a range of priors, such as skeptical and enthusiastic priors. A skeptical prior might be one in which the intervention is expected to have no effect, while an enthusiastic prior might be one where the effect is expected to be positive. A conclusion of a positive effect is strengthened if it is found even with a skeptical prior, and a conclusion of no effect is strengthened if it is found with an enthusiastic prior. For example, the Medical Trial Council Misonidazole Trials tested the sensitivity of the results to three prior distributions: an uninformative prior, a skeptical prior with an average effect of zero, and an enthusiastic prior with a positive average effect.[12] When the results showed the intervention was unlikely to be sufficiently successful even under the enthusiastic prior, the analysts were more comfortable concluding that the intervention should not be used.

Examples of prior distributions. Figure 1 shows several examples of prior distributions for a hypothetical program designed to increase earnings.

- The solid line is meant to represent an uninformative prior. Its weight is distributed evenly across the range presented in the figure. This type of prior

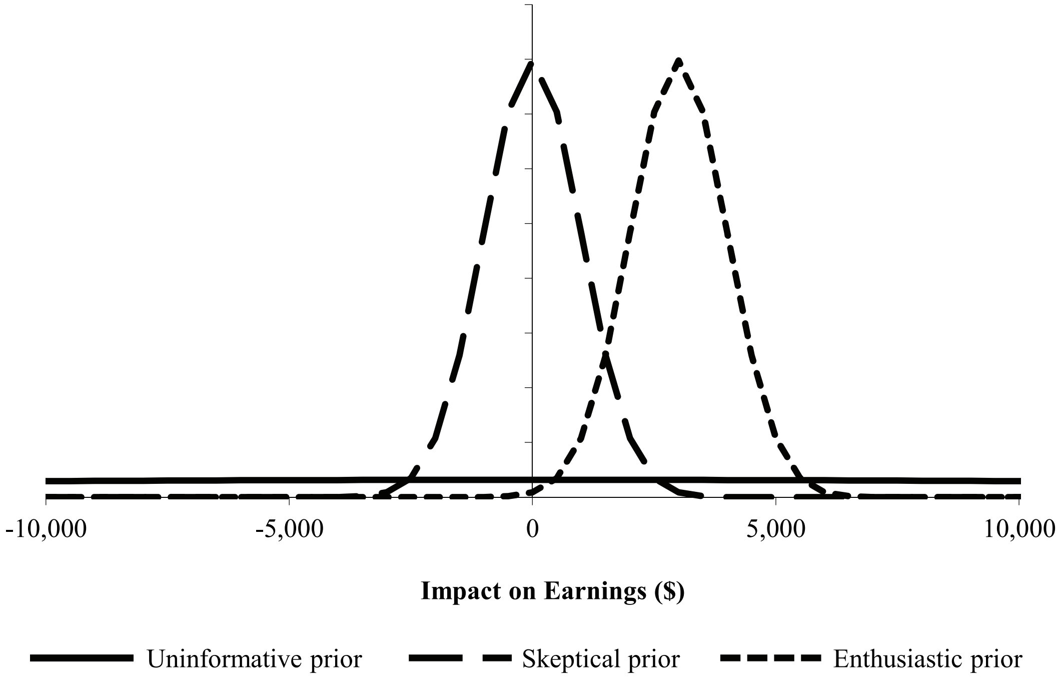
[8] Berk, Campbell, Klap, and Western (1992).
[9] Spiegelhalter, Freedman, and Parmar (1994).
[10] Vail, Hornbuckle, Spiegelhalter, and Thornton (2001).
[11] Spiegelhalter, Freedman, and Parmar (1994).
[12] Spiegelhalter, Freedman, and Parmar (1994).

Enhanced Services for the Hard-to-Employ Demonstration

Figure 1

Three Prior Distributions of the Effects on Earnings

Bayesian Reanalysis



might be used if there is no information on which to base a reasonable prior or as part of a sensitivity check.

- The long-dashed line represents a somewhat skeptical prior, indicating that the impact of the intervention is likely to be around zero. Compared with the uninformative prior, new data would have to provide greater evidence of a positive effect to lead to a positive assessment of the intervention.

- The short-dashed line represents a more enthusiastic prior. The prior indicates that the intervention is very likely to have an effect near $3,000 and that it is as likely to have an effect exceeding $3,000 as falling below that level. This prior is the most likely of the three priors to yield a conclusion of a positive program effect when combined with new evidence.

Collecting new evidence. The next step is to collect new evidence on the question being investigated — for example, by following a randomly assigned group of individuals.

Calculating the posterior distribution. The next step is to combine the prior with the new data to form a posterior distribution. For randomized trials of any size, the estimated effect will be approximately normally distributed. Let's say the program-comparison difference is m with a standard error of s. A normal prior with mean μ and variance σ^2 combined with these new data will form a posterior distribution that is normally distributed with a mean that is a weighted average of m and μ and a variance that is a weighted average of σ^2 and s^2. In particular, the mean of the posterior distribution will be

$$m' = \frac{m\sigma^2 + \mu s^2}{\sigma^2 + s^2}$$

The variance of the posterior distribution will be

$$s'^2 = \frac{\sigma^2 s^2}{\sigma^2 + s^2}$$

Thus, more precise priors will result in more precise posterior distributions with means closer to the prior mean, compared with less precise priors. Larger data collection efforts (such as larger randomized trials) will lead to more precise posterior distributions that look more like the new data rather than the prior.

This formula has an analog in classical meta-analytical methods. In pooling results across studies, the estimate with the lowest variance gives more weight to more precise estimates. This is exactly what is done above in the expression for the posterior mean. The mean from the prior distribution and from the new data are combined so that each receives a weight that is inversely proportional to its variance.

Interpreting the results. As noted earlier, Bayesian and classical methods interpret results differently. The classical approach yields information about the likelihood that a particular true impact would have generated the observed data. Suppose 200 individuals are randomly assigned to a training program where average program group earnings in the year after random assignment were $10,000 compared with $9,000 in the control group, with a standard error of $2,000. The classical approach would note that there is a 31 percent chance that a program with no real effect could have generated an impact of $1,000 or larger with this standard error. That is, the p-value of the null hypothesis of no effect is 0.31. The result would not be considered statistically significantly different from zero at conventional significance levels.

By contrast, Bayesian posteriors are interpreted as providing information on the distribution of the true impact, reflecting the analyst's uncertainty about the true effect.[13] Table 2 shows the mean and variances of the posterior distributions that result from the prior distributions shown in Figure 1 when combined with data from the 200-person experiment. The uninformative prior yields a posterior with a mean effect of $1,000 and a standard deviation of $2,000. Note that this is the same as the experimental difference of $1,000 with a standard error of $2,000 that is shown in the first column. The Bayesian posterior summarizes the analyst's uncertainty about the true effect by saying that the effect is centered at $1,000 with a standard deviation of $2,000. Inferences can be made about how much of the posterior distribution lies in some range. There is a 95 percent probability that the impact lies approximately between −$1,000 and $3,000, for example, and a 15.9 percent chance that the impact is greater than $3,000.

Because the uninformative prior does not influence the posterior distribution, inferences about the distribution of effects could be made in this case without formally invoking the Bayesian prior. That is, the experimental evidence provides an impact estimate with a mean difference of $1,000 and a standard error of $2,000. Since impact estimates are approximately normally distributed, this would suggest that there is a 31 percent chance that the program's effect is positive and a 15.9 percent probability that the program's effect is greater than $3,000. Although these types of inferences can be drawn from the experimental data alone, they represent a Bayesian approach to drawing inferences and are not, strictly speaking, consistent with the classical approach.

Table 2 also shows that the skeptical prior yields a distribution that is less favorable to the possibility that the program is successful and suggests that there is only a 0.1 percent chance that the impact is greater than $3,000. The enthusiastic prior yields a posterior distribution that is more favorable to the program and suggests that there is a 32.7 percent chance that the program's true effect is at least $3,000. While the result that one chooses to believe depends on which prior is closest to one's own belief, it is clear in this analysis how different beliefs affect the conclusions.

Why Use Bayesian Updating?

Any given impact estimate is a combination of true impact and sampling error and is therefore consistent with a range of true impacts. The Bayesian prior provides another piece of information by which to judge whether estimates are "real" or are likely to be outliers. If new data are consistent with the prior, the Bayesian posterior will lead to a firmer conclusion than the classical approach, just as a meta-analysis of several studies can lead to firmer conclusions

[13] Bolstad (2007).

Enhanced Services for the Hard-to-Employ Demonstration

Table 2

Bayesian Analysis of Hypothetical Earnings Impacts with Different Priors

Bayesian Reanalysis

		Summary of Bayesian Posterior Distributions		
Measure	Mean Difference from Experiment	Uninformative Prior	Skeptical Prior	Enthusiastic Prior
Mean ($)	1,000	1,000	200	2,600
Standard deviation or error ($)[a]	2,000	2,000	894	894
Probability impact is greater than $3,000 (%)	—	15.9	0.1	32.7

NOTE: [a]Results from a Bayesian analysis are summarized using the standard deviation. Results from a classical analysis are summarized using the standard error.

than any one study. This may help policymakers make decisions even with relatively small studies. For example, if there is substantial evidence of a program's effects for one population or in one site, a relatively small confirmation study may provide adequate evidence on which policymakers can base a decision to extend the program to a different subgroup or a different site. The Bayesian approach may provide an alternative to conducting another large evaluation designed to generate statistically significant impacts.

By contrast, a Bayesian analysis might call into question unusually large impacts from an intervention that was expected to have modest effects, especially if the pattern of effects does not conform to theoretical expectations. To give one example, MDRC's final report from the Minnesota Family Investment Program gained some attention for finding that the program increased marital stability by nearly 20 percentage points among two-parent welfare recipient families.[14] However, there was little precedence for such an effect from other studies of welfare reform policies, the program's effects on the direct targets of employment and welfare receipt were quite modest (with no change in employment and a 5 percent increase in income), and information on marriage was available for only a small sample of survey respondents. In this case, a Bayesian approach might have suggested that the actual effects were much smaller than indicated by the survey sample, and a later analysis with administrative records did find a much

[14] Miller et al. (2000).

more modest difference.[15] This is not a unique phenomenon: surprising findings are rarely replicated in later studies.[16]

Thus, a Bayesian approach formalizes what wise analysts do, to some extent, anyway. Positive experimental results are more persuasive when the treatment is founded on a strong theory. Positive results that are not theoretically based or larger than expected should be viewed with more skepticism.

Allowing inferences on distributions or ranges of impact estimates also presents some advantages to policymakers, especially when decisions depend not just on whether a program's effect is not zero, but whether it falls in a particular range. For example, a policymaker may want to adopt an approach that appears to be cost-effective compared with alternatives. The ability to make statements about the range of impacts may also be useful in conducting small studies. Consider the case of treating mothers with substance abuse problems through supportive housing, which provides stable housing along with treatment supports. Because supportive housing is substantially less expensive than providing substance abuse treatment in a residential facility, a state may want to use it as long as the evidence does not suggest that parents and children are worse off than with treatment in a residential facility. An evaluation conducted with 20 families in supportive housing and 20 families in residential treatment could provide inferences on the probability that supportive housing does not make families worse off if it were conducted from a Bayesian perspective. By contrast, a classical perspective would most likely show insignificant differences between the two approaches and therefore be uninformative.

As discussed above, the Bayesian approach is also viewed as a means of making sensitivity checks before drawing conclusions. If both a skeptical prior and an enthusiastic one yield positive conclusions, that should increase confidence that the intervention is effective. If, by contrast, a skeptical prior reverses the positive results of an experiment, a policymaker might want to collect more information before deciding whether to implement the intervention on a wider scale.

While Bayesian priors may introduce some subjectivity, Donald Berry provides an intriguing contrast to help explain why the Bayesian perspective can be a sensible one.[17] Berry describes how Anna Pierce used an experiment on one calf to draw a conclusion about the cause of a disease called "trembles." After a Shawnee woman told Pierce that white snakeroot was responsible for trembles, Pierce fed the plant to a calf. When the calf developed trembles, Pierce's strong prior — based on the Shawnee woman's claim — allowed her to infer that white

[15] Gennetian (2003).
[16] Ioannidis (2005).
[17] Berry (1996).

snakeroot was the culprit. Berry then describes Sam Carter's attempt to draw an inference from one observation that led to a more questionable conclusion. Observing that a drought ended in Los Angeles when two balloons were released, he concluded that releasing balloons could cause rain to fall. Carter's conclusion seems absurd because the prior that hot air balloons affect the weather seems questionable (and presumably could be disproved with some experimentation).

Examples of Bayesian Analyses in Random Assignment Studies

Bayesian updating has rarely been used as the primary approach for drawing inferences from social policy experiments, but it has been commonly used in other contexts, especially in medical research, where it has been used in a number of ways.

Estimating Mean Impacts

Bayesian updating has been used in assessing an intervention's likely effects. An example of Bayesian updating in social policy experiments is the spouse abuse experiment in Colorado Springs.[18] An earlier experiment conducted in Minneapolis randomized police officers responding to domestic abuse calls to do one of three things: arrest the batterer, ask the batterer to leave the premises, or try to restore order.[19] The experiment showed a sizable reduction in subsequent incidents of violence as a result of arrest, prompting a number of law enforcement agencies to use a similar approach. Because the Minneapolis study had a sample of only 314, the study was replicated in five locations: Charlotte, North Carolina; Colorado Springs, Colorado; Dade County, Florida; Milwaukee, Wisconsin; and Omaha, Nebraska. When data from Colorado Springs became available, results were already available from Omaha and Milwaukee.

Some researchers were concerned about a particular finding in the Omaha and Milwaukee replications:[20] violence went down among those who were employed at the time of random assignment but increased among others. Although data from Colorado Springs alone did not show the same pattern, a Bayesian analysis using a prior based on results in Omaha and Milwaukee suggested larger subgroup differences than the classical analysis of Colorado Springs had suggested. To test the sensitivity of their assumption about the prior distribution, the authors investigated a range of priors, from putting no weight on the previous studies to weighting data from the three other sites equally. The authors conclude that the evidence suggests that the intervention reduced spousal violence among suspects who were employed or in the military at the time of random assignment and increased it among others.

[18] Berk, Campbell, Klap, and Western (1992).
[19] Sherman and Berk (1984).
[20] Berk, Campbell, Klap, and Western (1992).

Although Bayesian updating has not often been used in social policy experiments, it has more frequently been used in experiments in medical research. For example, Bayesian updating has been used in several medical trials.[21] In the Medical Research Council Neutron Therapy Trial, even enthusiastic priors indicated that the effect of the new therapy was unlikely to be clinically worthwhile. In the Continuous Hyperfractionated Accelerated Radiotherapy Study in Head and Neck Cancer, divergent opinions of 10 experts were used to help shape expectations about the effects a new therapy. In the Levamisole and 5-fluorouracil in Bowel Cancer study, skeptical and enthusiastic priors were used in a sensitivity check of a new therapy's effect on mortality among cancer patients. Finally, the Medical Trial Council Misonidazole Trials used three prior distributions to test the sensitivity of results in a randomized trial of various therapies with 108 individuals with brain, head and neck, and cervical cancer. The results indicated that the treatment was unlikely to be effective even under the enthusiastic prior, which weighed against adopting the new therapy.

Comparing Subgroups or Sites

As the example from the Colorado Springs experiment suggests, Bayesian methods can be used in random assignment studies to analyze effects by subgroup. The usual starting point in Bayesian analyses of subgroup impacts is a random effects framework, which assumes that the true impacts for different subgroups can be characterized by a probability distribution. For example, each subgroup's true impact might be thought of as a random draw from a normal distribution. A Bayesian random effects analysis results in an estimate for a subgroup or site that is a weighted average of the subgroup or site impact and the pooled, overall impact.[22] This method results in individual subgroup estimates that are "pulled back from ordinary least squares estimates … towards the common mean."[23] The intuition behind this is that estimated subgroup impacts that are far from the mean probably include unusually large sampling errors. At one extreme, the analyst might attach no weight to the individual subgroups and use the overall impact as the estimated impact for each subgroup. At another extreme, the analyst could ignore the pooled mean and base subgroup results on data for only that subgroup. The middle option would put some weight on both the subgroup impact and the pooled impact. Which of these is chosen depends on how much previous information the analyst has for distinguishing among subgroups. For example, in the spouse abuse experiments, information on the two key subgroups was available from earlier experiments. In a multisite experiment of one program model, there might be little reason to expect larger effects for one site than another, which might lead the researcher to use a common prior for each site.

[21] Spiegelhalter, Freedman, and Parmar (1994).
[22] Gelman, Carlin, Stern, and Rubin (2004).
[23] Breslow (1990).

Overview of the Paper

The remainder of this paper uses Bayesian methods to reexamine results from the Enhanced Services for the Hard-to-Employ Demonstration and Evaluation Project. The Hard-to-Employ study included four sites, each of which had modest estimated effects that sometimes puzzled the research team. The main question addressed in this paper is whether a Bayesian approach tends to confirm or contradict the published results.

The next section reexamines results from a care management program provided via telephone to Rhode Island Medicaid recipients with depression. The classical analysis from that study found little effect on depression severity from encouraging individuals to seek and receive treatment for their depression. Results from the most similar previous efforts to use telephonic care management were not overly positive either, so that the Bayesian analysis confirms that the program is unlikely to have had a large effect.

Following the analysis of the Rhode Island program is a reexamination of the results from two transitional jobs programs, one operated in New York City for recently released prisoners and one operated in Philadelphia for hard-to-employ welfare recipients. In both cases, the classical analysis indicated that the programs had little effect on employment once participants' eligibility for transitional jobs had ended. However, impacts in the New York study were larger for the most recently released prisoners. In addition, some scattered long-term effects on recidivism in the criminal justice system were observed. The Bayesian analysis suggests that the New York program is likely to have had small positive effects on (that is, increased) employment and small negative effects on (that is, reduced) recidivism in the second year, but that impacts for the recently released group may not be as large as suggested in the study's original classical analysis.

In Philadelphia, the reanalysis shows how sensitive the results can be to how Bayesian priors are chosen. Previous studies of transitional jobs (subsidized, temporary jobs to help unemployed individuals make the transition to work) for welfare recipients are quite old, and two earlier studies included significant training components. Using the programs with training to form priors results in estimates that suggest that transitional jobs are very likely to have modest effects on employment and welfare receipt in the medium term. In contrast, when those programs are excluded, the reanalysis of Philadelphia's transitional jobs program confirms the evaluation's findings from the classical analysis that the program had little effect on unsubsidized employment and welfare receipt in the medium term.

In Kansas and Missouri, an enhanced version of Early Head Start — an early childhood education program for children from low-income families — found larger impacts at 12 months on a range of outcomes for families with pregnant mothers or with infants at the time of random assignment than for those with toddlers. This is consistent with findings from a national evaluation of Early Head Start, which found larger effects for families where the mother was pregnant than for those with infants. A Bayesian perspective might use priors for impacts on pregnant mothers and mothers of infants from the national evaluation to draw stronger inferences about the effects for families with infants in the Kansas-Missouri site. However, this analysis could not be conducted because data on outcomes where impacts were larger for families with infants in the Kansas-Missouri site were not collected in the national evaluation.

Results of the Bayesian analysis in Rhode Island, New York City, and Philadelphia generally confirm the published, classically derived findings that most impacts from these programs tend to be small. This is in part because results for the three sites are broadly consistent with similar studies that have been conducted, but in part because each of the sites included a relatively large sample. The Bayesian framework may be more informative when applied to smaller studies that might not be expected to provide statistically significant impact estimates on their own.

Rhode Island: Working toward Wellness

Although low-income individuals are disproportionately likely to suffer from depression, few receive treatment and even fewer stay with their treatment.[24] Untreated depression can negatively affect employment, job performance, and worker productivity.[25] The Working toward Wellness (WtW) study tested the effects of care management provided via telephone to depressed Medicaid recipients in Rhode Island to encourage them to seek treatment for their depression.[26] Master's-level clinicians — care managers — called individuals to encourage them to seek treatment, to help them find and make appointments with mental health professionals, to make sure they were keeping appointments and taking prescribed medications, to educate them about how depression would affect them and how treatment could help them, and to provide telephone counseling to individuals who were reluctant to seek treatment in the community. Although telephonic care management has been shown to be effective in treating depression with some populations,[27] WtW was the first study of the approach with low-income parents who are Medicaid recipients.

Results from the evaluation of WtW showed that program group members visited mental health professionals twice as often as did control group members during the first six months, although program and control group members were about equally likely to fill a prescription for antidepressants. Despite the increase in mental health visits, WtW did not significantly reduce

[24] Adelmann (2003); Kessler et al. (2003); Belsher and Costello (1988).

[25] Danziger et al. (2002).

[26] Kim, LeBlanc, and Michalopoulos (2009).

[27] Wang et al. (2004); Simon et al. (2004).

average depression scores for the full sample. It did reduce depression severity for a Hispanic subgroup, however — a group for which it also increased the use of antidepressants.

The mixed nature of the results through six months raises some questions about how to interpret the effects of WtW. The doubling of mental health visits was expected to lead to improvements in depression severity, and it is possible that a small but real reduction in depression severity fell short of standard levels of statistical significance. In addition, the larger effect for Hispanic sample members was intriguing and encouraging, but it was highly uncertain because it applied to a relatively small subgroup.

To further investigate results from WtW, this section reanalyzes results from the study by formally incorporating Bayesian priors. For example, evidence from previous studies suggests a link between care management and treatment for mental health, as well as a link between treatment for mental health and reductions in depression severity. When incorporated into a Bayesian framework, that earlier evidence provides information that can be useful in disentangling statistical noise from the effects of the intervention.

Forming Priors for Working toward Wellness

There is quite a large literature on the use of medications and psychotherapy to reduce depression. For forming priors for WtW, the analysis was limited to three random assignment studies that used telephonic care management to encourage individuals to seek mental health treatment and two studies that used in-person care management with more disadvantaged groups (to inform questions about the effectiveness of WtW for the Hispanic subgroup). These studies were reviewed by the WtW team for the main evaluation in consultation with Greg Simon, M.D., from Group Health Cooperative (GHC), as most relevant for the intervention that was used or the target population. The study team did not do a comprehensive search for relevant articles, as might be conducted in a meta-analysis. Nevertheless, these studies should be instructive for understanding how a Bayesian approach might have altered the study's conclusions.

The most directly relevant previous study is the Work Outcomes Research and Cost-Effectiveness Study (WORCS),[28] on which WtW was based. WORCS provided telephonic care management to active employees of large corporations who were covered by employer-sponsored health insurance but were not in active treatment for depression. WORCS found that telephonic care management increased the number of visits to health care providers to treat mental health conditions and modestly reduced depression severity, but did not increase the use of antidepressants.

[28] Wang et al. (2007).

Two other studies of telephonic care management were conducted by researchers at GHC. One study randomized 613 GHC patients who had recently been prescribed antidepressant medications to two programs to improve depression outcomes and to a control group.[29] One program provided feedback to doctors 8 and 16 weeks after medications were first prescribed. The second program included both feedback to doctors and telephonic care management to patients. Feedback alone did not affect patient outcomes, but feedback plus care management increased the use of appropriate doses of antidepressants and reduced depression severity. Neither intervention increased the number of mental health visits.

A second GHC study randomized individuals to three groups: usual primary care, telephonic care management, and telephonic care management plus telephonic psychotherapy.[30] Participants were GHC primary care patients beginning antidepressant treatment for depression. Compared with usual primary care, telephonic care management increased the number of mental health visits and the use of antidepressants, and it reduced depression severity.

Because the target population for WORCS is closest to WtW and because the WtW intervention is based on WORCS, that study may provide the best source of previous information about the likely effects of WtW. WORCS increased the number of mental health visits by 0.5 over the first six months, had no effect on the use of antidepressants, and reduced depression severity by 0.2 standard deviation.[31]

Although WORCS is the closest comparator to WtW, the GHC studies might also provide useful information about the likely effects of WtW. Each of these studies presented results from a follow-up at six months, and that time period was chosen for the current analysis. On average, the three earlier studies increased the number of mental health visits by 0.5 per person, increased the proportion taking an appropriate dose of antidepressants by 5.3 percentage points, and increased the likelihood that depression scores would be reduced by 50 percent or more by 8.8 percentage points. The two studies that reported depression severity scores reduced depression severity by 0.2 standard deviation on average.

These three studies focused on broad populations rather than low-income groups. Two other previous studies have shown that more intensive care management than was tested in WtW can increase mental health treatment and reduce depression levels for more disadvantaged families. Partners in Care found that intensive care management by nurses in primary care

[29] Simon, VonKorff, Rutter, and Wagner (2000).

[30] Simon et al. (2004).

[31] In Wang et al. (2007), mental health visits were classified as visits to any clinician — a primary care physician, a mental health specialist, or other clinician — for depression. The article did not report whether individuals had an appropriate dose of antidepressants, but it reports that the impact on use of any antidepressants was not statistically significant.

settings increased the use of appropriate levels of medications and reduced depression levels for a sample that was about one-third minority.[32] Results from Partners in Care were similar for minority and white sample members. A study called "WE Care" found that antidepressant medications administered by a primary care nurse practitioner reduced depression levels for a sample that was primarily Latina and African-American.[33] The two studies suggest that care management is effective for minority groups, although they do not indicate that the effects of care management are greater for minority group members than for others.

Results of the Bayesian Analysis for Working toward Wellness

For WtW, the reanalysis examines whether the intervention increased mental health treatment and reduced depression severity for the full sample, and whether it had larger effects for the Hispanic subgroup. As noted in the Introduction to this paper, some Bayesian statisticians have suggested testing the sensitivity of Bayesian results to a range of the priors. Therefore, several sets of priors were used: (1) an uninformative prior, (2) a prior based on the effects from WORCS, and (3) a prior based on mean results from WORCS plus the two GHC studies.

Full Sample Impacts

The first phase of the analysis examines the implications of a Bayesian perspective on the estimated impacts of WtW for the full sample. The analysis examines impacts on four outcomes: (1) the number of mental health visits, (2) whether the person received an adequate dosage of antidepressants, (3) average depression scores, and (4) whether an individual's depression score declined by 50 percent or more between baseline and the six-month survey. These outcomes were chosen because they were most often reported in the studies that were used to form priors. Results are summarized in Table 3.

Classical perspective. The first panel of Table 3 shows impact estimates for WtW from a classical perspective. According to this analysis, WtW increased the number of mental health visits from 1.5 for the control group to 2.6 for the program group. The difference of 1.1 visits has a standard error of 0.4. The p-value of 0.014 indicates that there is only a 1.4 percent chance that a program with zero true effect would have resulted in an estimated effect of 1.1 visits. Since impact estimates with p-values of 0.100 or less were considered significantly different from zero by the study team, this estimated effect is considered to be statistically significant.[34]

The remaining results in the first panel indicate that WtW did not significantly increase adequate dosages of antidepressants, did not significantly reduce average depression severity,

[32] Miranda et al. (2004); Wells et al. (2000).

[33] Miranda et al. (2003).

[34] Asterisks, a common convention to indicate statistical significance, are not used in the tables in this paper.

Enhanced Services for the Hard-to-Employ Demonstration

Table 3

Estimated Effects of Working toward Wellness at Six Months: Summary of Posterior Distribution Using Bayesian Analysis

Bayesian Reanalysis

Outcome	Program Group	Control Group	Difference (Impact)	Standard Deviation or Error[a]	P-Value	Probability of Positive Impact
Standard impact estimates[b]						
Number of mental health visits	2.6	1.5	1.1	0.4	0.014	
Took adequate dosage of antidepressants (%)	26.4	21.1	5.2	4.4	0.240	
Normalized depression score[c]	2.2	2.3	-0.1	0.1	0.340	
Depression improved by 50 percent (%)	17.8	19.5	-1.6	4.1	0.685	
Uninformative priors						
Number of mental health visits	2.6	1.5	1.1	0.4		99.4
Took adequate dosage of antidepressants (%)	26.2	21.0	5.3	4.4		88.3
Normalized depression score[c]	2.2	2.3	-0.1	0.1		17.0
Depression improved by 50 percent (%)	17.5	19.2	-1.7	4.1		34.2
Priors based on WORCS						
Number of mental health visits	2.5	1.6	0.9	0.4		99.4
Took adequate dosage of antidepressants (%)	23.5	23.7	-0.2	2.8		47.6
Normalized depression score[c]	2.2	2.4	-0.2	0.1		0.4
Depression improved by 50 percent (%)	19.4	17.4	2.0	2.4		79.9
Priors based on average of previous research						
Number of mental health visits	2.3	1.7	0.6	0.1		100.0
Took adequate dosage of antidepressants (%)	25.9	21.3	4.6	2.8		95.1
Normalized depression score[c]	2.2	2.4	-0.2	0.1		0.0
Depression improved by 50 percent (%)	19.8	16.8	3.0	2.4		89.9
Sample size (total = 370)	187	183				

(continued)

Table 3 (continued)

SOURCES: Wang et al. (2007), Simon et al. (2004), standard impact estimates from author calculations using Stata, and Bayesian results from author calculations using WinBUGS14.

NOTES: [a]Results from a Bayesian analysis are summarized using the standard deviation. Results from a classical analysis are summarized using the standard error.

[b]Results for the standard impact estimates differ from published results for Working toward Wellness because they are limited to 6-month survey respondents and use a different regression adjustment.

[c]Depression severity was measured using different standard instruments in the various sources, so scores are normalized to have a standard deviation of 1 within each study. A larger normalized score signifies more severe depression.

and did not significantly increase the proportion of individuals whose depression scores declined by 50 percent or more. In each case, the p-value is greater than 0.100.

Uninformative Bayesian prior. The second panel of Table 3 shows results using an uninformative Bayesian prior.[35] Several measures of the posterior distribution are presented for each prior. The first two columns show the regression-adjusted mean outcomes for the program and control groups. The third column shows the mean of the posterior distribution of the effect, while the fourth column shows the standard deviation of the distribution. The final column shows the probability that the estimated effect is greater than zero.

As noted earlier, the uninformative prior results in a posterior distribution that reflects the data collected for the experiment. In this case, the mean estimated impact on number of mental health visits is 1.1 using both approaches. Likewise, the posterior mean is similar to the estimated mean effect using the classical approach for use of antidepressants (5.3 versus 5.2 percentage points), for depression severity (−0.1 for both approaches), and for whether depression severity was improved by at least 50 percent (−1.7 percentage points for both approaches). In addition, the standard errors of the classical results are similar to the standard deviations of the posterior means from the Bayesian analysis with uninformative priors. For example, the standard error of the standard impact estimate for number of mental health visits is 0.44, and the standard deviation of the posterior distribution from the Bayesian analysis is 0.44. Differences in the results are due to the fact that the Bayesian results are generated using an iterative method, while the classical results are not.

[35] Because many Bayesian statistical analyses do not have simple solutions, the Bayesian results were obtained using Monte Carlo Markov chain methods within a computer package called WinBUGS. Results are presented as the mean and standard deviation of the Bayesian posterior estimates, summarized across 20,000 iterations.

As discussed in the Introduction, the two statistical approaches interpret the results differently. The classical approach draws inferences about the probability that a program with zero effect could have generated the estimates, and this probability is summarized by the p-value. By contrast, the Bayesian posterior is interpreted as indicating the distribution of the true impact, which is not known. That allows for inferences on any aspect of the distribution of effects, such as the probability that the impact is positive. The last column of Table 3 shows the estimated probability that the impacts are positive.

Using the uninformative prior, there is a 99.4 percent chance that the effect of WtW on the number of mental health visits is positive. This is roughly consistent with the classical finding that there is a 1.4 percent chance that the effect was zero, but it is important to recognize that these are two different concepts. One is making a statement about whether a program with a true effect of zero could have generated the observed result, while the other is making a statement about the probability that the true effect is positive.

The difference in interpretations is most striking in drawing inferences about the effect of WtW on the use of appropriate doses of antidepressants. The classical approach shows a p-value of 0.240 — not significantly different from zero — while the Bayesian approach indicates that 88.3 percent of posterior distribution is above zero; that is, there is an 88.3 percent chance that the impact is positive. The results are approximately the same, but one interpretation suggests that the impacts are indistinguishable from zero using conventional levels of statistical significance while the other suggests that the impact is very likely to be positive. Likewise, the classical approach produces a p-value of 0.340 for the impact on depression severity, while the Bayesian posterior is negative 83.0 percent (100 minus 17.0) of the time. One analyst would conclude that there is not enough evidence to suggest that the effect is not zero, but the other analyst would conclude that the program is very likely to have reduced depression severity.

Prior based on WORCS. The third panel of Table 3 presents results of the Bayesian analysis using priors developed from WORCS. As noted earlier, WORCS found an increase in the number of mental health visits, a reduction in depression severity, and an effect on using antidepressants that was not statistically significant. Using WORCS as a prior should consequently reinforce the WtW findings regarding mental health treatment and strengthen conclusions about its effects on depression severity.

Because WORCS had a slightly smaller effect than WtW did on the number of mental health visits, using this prior reduces the mean of the posterior distribution of the effect on mental health visits to 0.9. Nevertheless, the evidence is quite strong that WtW increased the number of mental health visits: there is a 99.4 percent probability that the effect is positive. This is because the small positive effects from WORCS and WtW combine to increase confidence in the result.

Because the program group in WORCS was actually less likely to take antidepressants than the control group, using this prior suggests that WtW did not affect antidepressant use (rather than having a small but statistically insignificant effect). The mean estimate is now very close to zero.

For depression severity, priors from WORCS lead to somewhat more positive conclusions. The mean impacts are a reduction of 0.2 standard deviation in depression severity and a 2.0 percentage point increase in depression improving by at least 50 percent. There is a 99.6 (100 − 0.4) percent chance that WtW reduced depression severity and a 79.9 percent chance that WtW increased the proportion of sample members whose depression score declined by at least half.

Prior based on three earlier studies. The fourth panel of Table 3 shows results when the Bayesian prior is based on all three telephonic care management studies described earlier. The table shows that including the positive results from the two GHC studies leads to stronger conclusions that the program was effective. Compared with the classical results or results with the WORCS prior, these results provide greater confidence that WtW had positive effects. There is a 100.0 percent probability that the program increased the number of mental health visits, an 95.1 percent chance that it increased appropriate use of antidepressants, a nearly 100 percent probability that it reduced depression severity, and an 89.9 percent chance that it increased the proportion of sample members whose depression score declined by at least half. By reducing the degree of uncertainty, however, the results also suggest it is unlikely that WtW had sizable effects. For example, 95 percent of the posterior distribution for the reduction in depression severity falls between 0.01 and 0.3 standard deviations, which would be small to modest reductions using Cohen's method of classifying effect sizes.[36]

Subgroup Differences

One of the intriguing findings from WtW is its short-term effectiveness for Hispanic sample members. Although Partners in Care and WE Care also found significant effects for minority sample members, the evidence that impacts will be larger for Hispanic sample members is unclear. Partners in Care found similar effects for Hispanic and other sample members,[37] while WE Care enrolled primarily minority sample members.

This section presents a relatively simple Bayesian reanalysis of the subgroup impacts. Because there is little evidence on the effects of telephonic care management for various ethnic groups, one prior is used for both ethnic subgroups. The subgroup results are, in essence, an average of the prior and the new data, pulling the subgroup estimates back toward that common

[36] Cohen (1988).

[37] Miranda et al. (2006).

prior. Three such priors are tested: (1) an uninformative prior, (2) a prior based on WORCS, and (3) a prior based on the three telephonic care management interventions. Results are presented in Table 4.

The first panel in Table 4 shows regression-adjusted impact estimates from a classical perspective. The estimates are statistically significant for three of the four outcomes for the Hispanic subgroup but for none of the outcomes for the non-Hispanic subgroup.[38] In addition, estimated effects on mental health visits are significantly larger for the Hispanic subgroup.

Using priors based on WORCS, differences across the subgroups are much smaller (third panel of Table 4). This is in part because results for the Hispanic subgroup in WtW are outliers compared with results from WORCS. For example, the estimated mean effect on number of mental health visits is 1.1 for the Hispanic subgroup and 0.6 for the non-Hispanic subgroup (compared with 0.5 from WORCS). Because of the smaller difference between the two subgroups under the Bayesian posterior distribution, there is only a 76.7 percent chance that the effect is larger for Hispanic sample members. The evidence that impacts were larger for Hispanic sample members is also weaker for the other three outcomes, compared with the classical analysis.

Using priors based on the mean of all three previous studies, differences in estimated effects on mental health visits are even smaller (fourth panel of Table 4). Now, the estimated mean effect on number of visits to mental health professionals is 0.6 for the Hispanic subgroup and 0.5 for the non-Hispanic subgroup, and there is only a 56.7 percent probability that the effect was greater for Hispanic sample members. In part, this is because the two GHC studies included individuals who were already in treatment for their depression, and there may have been less room for improvement in that outcome.

The results by subgroup highlight the importance of priors in deriving subgroup impacts. One case lets the subgroup estimates speak for themselves (as in the classical analysis) and shows generally larger effects for Hispanic sample members. Using the same prior for each subgroup brings the estimates closer together and calls into question whether there was a real difference. This is consistent with findings from Partners in Care that showed similar impacts for minority and non-minority sample members.

[38] Statistical significance of the subgroup estimates is not shown in the table, but impact estimates are statistically significant at the 10 percent level or better when the ratio of the impact estimate to its standard error is 1.65 or greater in absolute value. This criterion is met for estimates of the impacts on number of mental health visits, adequate dosage of antidepressants, and normalized depression score for Hispanic sample members.

Table 4

Estimated Effects of Working toward Wellness at Six Months for Hispanic and Non-Hispanic Subgroups: Summary of Posterior Distribution Using Bayesian Analysis

Bayesian Reanalysis

	Hispanic Subgroup		Non-Hispanic Subgroup		Subgroup Difference	
Outcome	Impact Estimate	Standard Deviation or Standard Error[a]	Impact Estimate	Standard Deviation or Standard Error[a]	P-value	Probability Hispanic Is Greater
Standard impact estimates[b]						
Number of mental health visits	2.0	0.8	0.7	0.5	0.181	
Took adequate dosage of antidepressants (%)	19.5	8.1	-0.6	5.3	0.039	
Normalized depression score[c]	-0.3	0.2	0.0	0.1	0.130	
Depression improved by 50 percent (%)	6.1	7.5	-4.5	4.9	0.232	
Uninformative priors						
Number of mental health visits	2.0	0.8	0.7	0.5		90.8
Took adequate dosage of antidepressants (%)	18.9	7.3	-0.6	5.4		98.4
Normalized depression score[c]	-0.3	0.2	0.0	0.1		6.4
Depression improved by 50 percent (%)	5.6	6.4	-4.5	5.0		89.4
Priors based on WORCS						
Number of mental health visits	1.1	0.5	0.6	0.4		76.7
Took adequate dosage of antidepressants (%)	0.0	3.1	-2.8	3.0		73.9
Normalized depression score[c]	-0.2	0.1	-0.1	0.1		19.2
Depression improved by 50 percent (%)	3.6	2.4	1.8	2.7		68.9
Priors based on average of previous research						
Number of mental health visits	0.6	0.2	0.5	0.2		56.7
Took adequate dosage of antidepressants (%)	6.2	3.2	2.7	3.1		78.3

(continued)

Table 4 (continued)

	Hispanic Subgroup		Non-Hispanic Subgroup		Subgroup Difference	
Outcome	Impact Estimate	Standard Deviation or Standard Error[a]	Impact Estimate	Standard Deviation or Standard Error[a]	P-value	Probability Hispanic Is Greater
Normalized depression score[c]	-0.2	0.1	-0.2	0.1		22.3
Depression improved by 50 percent (%)	4.6	2.4	2.9	2.7		68.7
Sample size (total = 6,358)	110		260			

SOURCE: Standard impact estimated from author calculations using Stata and Bayesian results from author calculations using WinBUGS14.

NOTE: [a]Results from a Bayesian analysis are summarized using the standard deviation. Results from a classical analysis are summarized using the standard error.

[b]Results for the classical analysis differ from published results for Working toward Wellness because they are limited to 6-month survey respondents and use a different regression adjustment.

[c]Depression severity was measured using different standard instruments in the various sources, so scores are normalized to have a

Working toward Wellness: Discussion

The early findings from WtW were somewhat disappointing in light of the success of previous telephonic care management interventions. A Bayesian perspective casts a more positive light on the results. This occurred in two ways. First, the positive findings from the previous studies increased the confidence that the effects of WtW are likely to be positive. For example, the consistent effects of previous care management programs on depression severity suggest that WtW probably modestly reduced depression severity.

Second, the Bayesian focus on distributions rather than statistical significance tests provides a more positive interpretation of the results. Even though estimated effects on antidepressant use and depression severity were not statistically significant by conventional standards, the preponderance of the evidence from the Bayesian analysis suggests that WtW is very likely to have increased the use of antidepressants and reduced depression severity.

New York and Philadelphia: Transitional Jobs

As explained earlier, transitional jobs are a form of subsidized employment that are typically targeted at hard-to-employ individuals and designed to help them make the transition to unsubsidized, or "regular," jobs.[39] In comparison with other types of subsidized work, transitional jobs programs typically provide services such as job search assistance and job placement to help individuals move to unsubsidized employment. The Hard-to-Employ evaluation studied transitional jobs programs in two sites: the New York Center for Employment Opportunities (CEO) program for recently released prisoners and the Philadelphia Transitional Work Corporation (TWC) program for hard-to-employ welfare recipients.[40]

The evaluation found that CEO had little effect on employment once the transitional jobs were ended. However, the program did reduce recidivism, a result that puzzled the research team to some extent because employment was initially thought to be the avenue to reducing involvement with the criminal justice system.[41] CEO's effects were concentrated in a group of

[39] Bloom (2010).

[40] Welfare recipients are recipients of Temporary Assistance for Needy Families (TANF).

[41] Transitional jobs programs for people coming out of prison are based in part on the observation that those who are able to find jobs are less likely to commit additional crimes and return to prison. However, most of the evidence on this question comes from nonexperimental studies in which the ability to find a job and the ability to remain out of the criminal justice system may be related to some other factors, such as ties to the community or family members, or motivation (Wilson, Gallagher, and MacKenzie, 2000; Bushway and Reuter, 2002). In addition, the literature suggests that employment must be stable if it is to affect recidivism, and transitional jobs were intended to last only a few months in the Transitional Jobs Reentry Demonstration (TJRD, discussed below) and CEO (Laub and Sampson, 2001). Perhaps for this reason, an examination of

(continued)

individuals who had been released from prison no more than 90 days prior to random assignment. Although the effects were larger for this subgroup, the impact estimates contain considerable uncertainty because the subgroup represents less than half of the study sample. For CEO, the Bayesian analysis explores the robustness of these findings. Overall, the analysis reinforces the finding that the program is likely to have increased employment and reduced recidivism. However, the magnitude of those effects for the recently released subgroup is probably smaller than suggested by classical analysis of the evaluation data.

The estimated effects of TWC on employment faded quickly after the transitional jobs ended. Despite the lack of a long-term effect on employment, TWC significantly reduced welfare receipt during the second year but not the third year. The Bayesian analysis addresses two questions. First, is the reduction in welfare receipt in the second year likely to be a true effect of the program? Second, are the small longer-term effects consistent with findings from other transitional work programs for welfare recipients, or is it more likely that TWC had some longer-term effects? The Bayesian results are somewhat sensitive to how priors are formed. That is because earlier studies of transitional jobs vary substantially in their effects, possibly because they took place over a relatively long period of time during which the welfare system was changing, or because they involved more education and training opportunities than were offered in in TWC.

Because the two transitional jobs programs targeted very different groups, results for the CEO and TWC programs are described in separate sections below. Each section describes some of the other studies that might shed light on the expected effects of the transitional jobs program studied in the Hard-to-Employ evaluation and then discusses the results of the Bayesian analysis for that program. In each case, priors were developed based on studies described by Bloom.[42]

New York: Center for Employment Opportunities

CEO provided recently released prisoners with several sets of activities: (1) a four-day Life Skills class focusing on job readiness, (2) a minimum-wage transitional job for four days per week at one of 30 to 40 public agencies administered through the Neighborhood Works Project, and (3) one day each week for meeting with job coaches or participating in other activities such as a fatherhood program.[43] Control group members were provided a version of the Life Skills class for one and a half days and were given access to a room that provided resources to help them look for employment.

impacts on employment and recidivism in TJRD and the National Supported Work Demonstration (discussed below) found little relationship between the two sets of effects.

[42] Bloom (2010).

[43] Redcross et al. (2009).

Participation in program services was fairly high. Nearly four out of five program group members completed the Life Skills class, and nearly 72 percent worked in a transitional job for an average of 8.3 weeks over a four-month period. In addition, nearly 60 percent met with a job coach and nearly 60 percent met with a job developer.[44] Perhaps because they were offered services such as the shorter Life Skills course, control group members reported being as likely as program group members to participate in activities such as job search and education, and they were half as likely as program group members to have worked in a transitional or subsidized job.[45]

Related Studies

For CEO, Bloom's review points to two other relevant studies.[46] The Transitional Jobs Reentry Demonstration (TJRD), which was conducted by MDRC and used similar outcome measures as CEO, took place in four cities: Chicago, Detroit, St. Paul, and Milwaukee.[47] Like CEO, the evaluation randomized former prisoners to a transitional jobs program or to a control group that was offered job search assistance.

In many respects, the TJRD transitional jobs programs were similar to the CEO transitional jobs program. Both TJRD and CEO provided participants with temporary, minimum-wage jobs that offered 30 to 40 hours of paid work each week. As in CEO, the TJRD transitional jobs aimed to identify and address behavior or performance issues that emerged at the work sites. TJRD also helped participants look for unsubsidized jobs.

Several differences might have produced different results from the two studies and across sites in TJRD. First, different organizations ran the TJRD transitional jobs and job search programs in Detroit, Milwaukee, and St. Paul, while one organization provided both program and control group services in the CEO evaluation. In addition, in TJRD, the Milwaukee and St. Paul transitional jobs programs offered employment bonuses, while the other two sites did not. While CEO began offering employment bonuses partway through that study, the use of such bonuses was more prominently described in the one-year report on TJRD, suggesting that they may have played a more important role in that study.[48] Since financial incentives have been found to encourage employment, this might have led to larger effects than the transitional jobs alone, and the study team has found some evidence of such effects.[49]

[44] Bloom, Redcross, Zweig, and Azurdia (2007).
[45] Redcross et al. (2009).
[46] Bloom (2010).
[47] Redcross et al. (2010).
[48] Redcross et al. (2010).
[49] Redcross et al. (2010).

TJRD targeted men who had been released from prison within 90 days, so its results might be especially relevant for understanding impacts for the recently released group in CEO. The sample includes just over 1,800 men. As in CEO, participation rates were high: 85 percent of the program group worked in a transitional job for an average of 53 days over four months.

A second relevant study is the 10-site National Supported Work (NSW) Demonstration, which generally provided 12 months of paid work experience that was supposed to gradually increase the expectations at the workplace until they approximated the conditions of an unsubsidized job.[50] Increased expectations might have included increased productivity demands, stricter attendance requirements, assignment to more complex tasks, or reductions in supervisory oversight. As in CEO, transitional jobs programs in NSW were run by nonprofit organizations, and individuals were assigned to small work crews. Transitional jobs could be with private-sector employers, including construction and small manufacturing industries, or in the public sector, such as working in public parks. As in CEO, time was available for support services such as skill training and job readiness and placement activities. However, only about 6 percent of paid time was used for these activities, whereas CEO set aside one day per week for job search and other activities.

NSW included two target groups that might be relevant to CEO: ex-prisoners and ex-addicts. The ex-prisoner group included individuals who had been incarcerated no more than six months before random assignment while the ex-addict group included men who had been in drug treatment in the preceding six months. Because results were similar for the two target groups, only results for the ex-prisoner group were used in forming priors for CEO. As in CEO, participation rates were high — attendance rates were more than 80 percent.

Table 5 summarizes results across the two studies. Pooling across the two studies, the estimated effects on quarterly employment declined from 40.3 percentage points near the beginning of the transitional jobs program to less than 2 percentage points in Year 2 (shown in Quarters 7 and 8).[51] Although the estimated effects for NSW increased again after Quarter 8, information in the third year was available for only about 300 of the 1,500 ex-prisoners, and those impacts are not statistically significantly different from zero using conventional statistical methods.

Neither TJRD nor NSW substantially reduced recidivism. For example, across the four TJRD sites and the NSW ex-prisoner group, estimated effects on arrests in Year 1 ranged from an increase of 1.6 to an increase of 5.4 percentage points, but none of these effects was statistically significant. Pooled results were generally close to zero.

[50] Manpower Demonstration Research Corporation Board of Directors (1980).

[51] The average was calculated by weighting each estimate by the inverse of estimated variance.

Enhanced Services for the Hard-to-Employ Demonstration

Table 5

Impacts on Employment and Recidivism in Transitional Jobs Programs for Former Prisoners

Bayesian Reanalysis

	Chicago TJRD		Detroit TJRD		St. Paul TJRD		Milwaukee TJRD		NSW Former Prisoners		Pooled	
Outcome (%)	Impact Estimate	Standard Error	Impact Estimate	Standard Error	Impact Estimate	Standard Error	Impact Estimate	Standard Error	Impact Estimate	Standard Error	Impact Estimate	Standard Error
Employment												
Ever employed												
Quarter 1	47.2	4.8	63.7	4.1	42.7	3.6	18.0	4.5	35.9	2.4	40.3	1.6
Quarter 2	31.5	5.1	54.6	4.4	27.0	4.0	27.3	4.3	23.2	2.5	29.7	1.7
Quarter 3	13.3	5.1	12.0	4.6	12.8	4.4	17.2	4.5	12.1	2.6	13.1	1.7
Quarter 4	3.3	4.9	2.9	4.3	5.4	4.4	7.6	4.3	4.1	2.6	4.5	1.7
Quarter 5	7.1	4.6	2.2	4.2	3.4	4.4	3.8	4.2	1.4	2.6	2.9	1.7
Quarter 6	4.9	4.5	0.8	4.0	5.7	4.3	2.9	4.1	0.8	3.1	2.6	1.7
Quarter 7	3.3	4.2	2.0	3.9	-1.5	4.0	4.3	4.0	1.5	3.1	1.9	1.7
Quarter 8	5.7	4.4	-2.3	3.8	1.1	3.9	1.7	3.8	-0.9	3.1	0.6	1.7
Quarter 9	7.3	4.3	1.7	3.9	-2.6	3.8	1.7	3.9	4.2	5.7	2.0	1.9
Quarter 10	4.5	4.2	1.2	3.9	-3.3	3.5	3.1	3.7	4.0	5.7	1.4	1.8
Recidivism												
Ever arrested												
Year 1	1.6	5.1	5.4	4.7	2.0	4.3	3.2	4.1	1.9	2.4	2.5	1.7
Year 2	-5.1	5.3	0.0	4.6	7.4	4.3	-3.5	3.8	2.4	2.6	0.9	1.7
Ever convicted												
Year 1	2.9	3.8	-3.9	3.6	3.4	3.1	-1.1	3.0	0.3	1.9	0.3	1.2
Year 2	-4.1	4.5	5.7	3.5	1.6	3.5	2.3	3.0	1.8	9.8	2.0	1.7

(continued)

Table 5 (continued)

	Chicago TJRD		Detroit TJRD		St. Paul TJRD		Milwaukee TJRD		NSW Former Prisoners		Pooled	
Outcome (%)	Impact Estimate	Standard Error	Impact Estimate	Standard Error	Impact Estimate	Standard Error	Impact Estimate	Standard Error	Impact Estimate	Standard Error	Impact Estimate	Standard Error
Ever incarcerated												
Year 1	2.5	4.3	3.0	3.4	-8.7	4.4	-1.4	4.5	0.5	2.3	-0.1	1.5
Year 2	-7.4	4.5	7.9	3.4	-3.4	3.9	-6.7	4.4	0.9	3.0	-0.4	1.6
Sample size	374		388		506		506		1,497[a]		1774	

SOURCES: Jacobs (2012) for Transitional Jobs Reentry Demonstration (TJRD) estimates and MDRC Board of Directors (1980) for National Supported Work Demonstration (NSW) estimates.

NOTES: Because transitional jobs in NSW lasted about three months more on average than in CEO, the first quarter of results from NSW were not used in forming the prior distribution.

[a] This sample size is for NSW quarters 1-5. Sample size for quarters 6-8 is 995, and for quarters 7-11 is 302.

Results of the Bayesian Analysis for CEO

Employment. Table 6 presents two estimates of CEO's effects on employment: (1) results using a classical approach based on the analysis conducted for the published CEO reports,[52] and (2) Bayesian results using mean impacts from TJRD and NSW to form a prior.

For the full sample, the two approaches provide generally similar conclusions. Using the classical approach, employment impacts for the full sample are not statistically significant after Quarter 3. For example, in Quarter 9, the impact estimate has a p-value of 0.113. The Bayesian posterior also suggests that employment impacts declined substantially after Quarter 3, although it provides a slightly more positive interpretation of the results. In Quarter 9, for example, the Bayesian results indicate that there is a 95.9 percent chance that CEO's effect on employment was positive. Although the classical perspective cannot reject the null hypothesis of zero effect in any quarter after Quarter 3, in every quarter the Bayesian perspective suggests that there is an 84 percent or greater chance that the effect is positive.

The difference partly reflects the use of a two-tailed t-test of statistical significance in the classical approach. Even with a one-tailed test, however, the classical framework would have yielded an estimate that was not statistically significant. A more important difference is in the interpretation of results. The classical approach examines the probability that CEO would have generated this estimated effect if it had a true effect of zero. The Bayesian approach assesses the probability that the effects are positive and suggests that they are probably small but very likely to have been positive. It is important not to overstate the differences between the two approaches. After Quarter 3, neither approach suggests CEO's effects on employment were more than small.

Because both TJRD and NSW targeted individuals who had been released from prison within three months, they provide a good basis for forming a prior for the recently released subgroup in the CEO study. Because the other studies found slightly less positive impacts, the Bayesian analysis produces somewhat weaker results than the classical analysis. Using the classical approach, the average quarterly employment rate from Quarter 5 on is 7.6 percentage points, while the mean of the Bayesian posteriors is only 2.6 percentage points.

Recidivism. Table 7 presents similar results for percentage arrested, percentage convicted of any crime, and percentage incarcerated, for Year 1 and Year 2. As above, results are presented using a classical approach and a Bayesian prior based on mean impacts from TJRD and NSW.

[52] Redcross et al. (2009); Redcross, Millenky, Rudd, and Levshin (2012).

Enhanced Services for the Hard-to-Employ Demonstration

Table 6

Estimated Impacts on Employment, Center for Employment Opportunities: Summary of Classical and Bayesian Analyses

Bayesian Reanalysis

	Classical Analysis			Bayesian Analysis		
Ever Employed (%)	Impact Estimate	Standard Error	P-Value	Impact Estimate	Standard Deviation	Probability of Positive Impact
Full sample						
Quarter 1	39.9	2.9	0.000	40.2	1.4	100.0
Quarter 2	24.3	3.1	0.000	28.5	1.5	100.0
Quarter 3	6.7	3.2	0.035	11.6	1.5	100.0
Quarter 4	1.4	3.1	0.637	3.8	1.5	99.5
Quarter 5	0.6	2.9	0.834	2.4	1.4	94.9
Quarter 6	1.7	2.9	0.561	2.4	1.5	94.4
Quarter 7	3.9	3.0	0.187	2.4	1.5	94.8
Quarter 8	4.1	2.9	0.168	1.5	1.5	84.4
Quarter 9	4.7	2.9	0.113	2.7	1.6	95.9
Quarter 10	2.9	2.9	0.308	1.8	1.5	88.1
Sample size	973					
Recently released subgroup						
Quarter 1	42.9	4.8	0.000	40.5	1.5	100.0
Quarter 2	34.3	5.0	0.000	30.2	1.6	100.0
Quarter 3	13.5	5.1	0.009	13.1	1.6	100.0
Quarter 4	8.1	5.0	0.105	4.9	1.6	99.9
Quarter 5	7.9	4.9	0.106	3.4	1.6	98.6
Quarter 6	4.9	4.8	0.304	2.9	1.6	96.1
Quarter 7	11.1	4.8	0.022	2.9	1.6	96.5
Quarter 8	8.0	4.7	0.086	1.5	1.6	82.5
Quarter 9	9.8	4.7	0.040	3.0	1.7	95.8
Quarter 10	4.1	4.6	0.371	1.7	1.7	84.8
Sample size	383					

SOURCES: Redcross, Millenky, Rudd, and Levshin (2012) and Bayesian results from author calculations using mean impacts from the Transitional Jobs Reentry Demonstration and the National Supported Work Demonstration.

For the full sample, CEO's effects on recidivism were not statistically significant using the classical approach in Year 1, and the Bayesian perspective confirms that CEO's impacts on recidivism were likely small in the first year. The two analyses show greater disagreement in Year 2. Results from CEO itself show a significant reduction in convictions of 6.3 percentage

Enhanced Services for the Hard-to-Employ Demonstration

Table 7

Estimated Impacts on Arrests, Convictions, and Incarceration, Center for Employment Opportunities: Summary of Posterior Distributions Using Bayesian Analysis

Bayesian Reanalysis

	Classical Analysis			Bayesian Results		
Outcome (%)	Impact Estimate	Standard Error	P-Value	Impact Estimate	Standard Deviation	Probability of Reduced Recidivism
Full sample						
Year 1						
Ever arrested	-1.7	2.7	0.532	1.4	1.4	16.4
Ever convicted	-2.0	2.4	0.399	-0.2	1.1	55.9
Ever incarcerated	-4.9	3.1	0.117	-1.0	1.4	77.7
Year 2						
Ever arrested	-5.3	2.8	0.056	-0.7	1.4	69.4
Ever convicted	-6.3	2.7	0.019	-0.4	1.5	60.2
Ever incarcerated	-4.3	3.1	0.174	-1.2	1.5	80.4
Sample size	977					
Recently released subgroup						
Year 1						
Ever arrested	-3.9	4.3	0.365	1.7	1.5	13.5
Ever convicted	-5.5	3.7	0.136	-0.3	1.2	58.8
Ever incarcerated	-10.2	5.0	0.044	-1.0	1.5	74.7
Year 2						
Ever arrested	-6.9	4.8	0.153	0.1	1.6	48.4
Ever convicted	-6.6	4.7	0.155	1.0	1.6	26.5
Ever incarcerated	-7.6	5.1	0.138	-1.1	1.6	75.2
Sample size	385					

SOURCES: Redcross, Millenky, Rudd, and Levshin (2012) and Bayesian results from author calculations using mean impacts from the Transitional Jobs Reentry Demonstration and the National Supported Work Demonstration.

points and a significant reduction in arrests of 5.3 percentage points. Because CEO's impacts on recidivism in Year 2 were much larger than the impacts in NSW or TJRD — which had essentially no effect on recidivism on average — the Bayesian perspective presents a much less positive picture. For example, it indicates that the effect on arrests has a mean reduction of only 0.7 percentage point and a 69.4 percent probability that arrests were reduced, and it indicates that the effect on convictions has a mean reduction of only 0.4 percentage point and the probability that the impact was negative is only 60.2 percent.

Results for the recently released subgroup generally suggest larger effects than for the full sample, but the smaller sample for this subgroup produces similar conclusions about whether CEO is likely to have reduced recidivism. Although the classical analysis indicates that CEO had a larger effect on all measures of recidivism for the recently released subgroup than for the full sample, the estimates are significantly different from zero for only one of the six estimates (incarceration in Year 1). The Bayesian posterior produces much smaller mean estimated effects and suggests that CEO probably did not reduce recidivism much for this subgroup. While the Bayesian results should not necessarily reverse the study team's conclusions about CEO, it does give room for skepticism about the results unless there is a convincing explanation for differences across the studies or CEO's effects can be replicated.

Philadelphia: Transitional Work Corporation

Welfare recipients were eligible for the TWC transitional jobs program if they had received TANF for at least 12 months or did not have a high school diploma or the equivalent, and if they were not currently employed or participating in work activities.[53] TWC's program began with a two-week orientation that taught job-readiness skills, followed by transitional employment, usually with a government or nonprofit agency. Individuals were employed by TWC, which paid the minimum wage for up to six months. Recipients were required to work 25 hours per week and to participate in 10 hours of professional development activities, such as job search and job-readiness instruction, as well as preparation for the General Educational Development exam and other classes. During the transitional work period, TWC staff worked with participants to find permanent, unsubsidized jobs for them. TWC also provided job retention services and bonus payments to participants for six to nine months after they were placed in a permanent job.

Participation in transitional jobs was lower than for CEO: 62 percent of those who were assigned to TWC completed the two-week orientation and half worked in a transitional job, working in one for 30 days on average. At the same time, control group members were also likely to receive some employment-related services, even though they were exempt from welfare-to-work activities. For example, although 76.1 percent of the TWC group participated in job search activities, 54.8 percent of the control group also did. Likewise, 33.9 and 38.1 percent of the two groups reported participating in education and training activities, and 18.3 percent of the control group reported some unpaid work.[54]

[53] Bloom et al. (2009).

[54] Bloom et al. (2009).

Related Studies

For welfare recipients, the Bloom review includes a larger number of potentially relevant studies, but all are fairly old.[55] In addition, a number of early tests of subsidized employment for welfare recipients had fairly low participation in subsidized jobs, making their results of questionable utility in understanding the effects of the TWC program.

As for former prisoners, NSW also included results for welfare recipients. As noted above, the program provided 12 to 18 months of supported employment that was designed to gradually introduce individuals to unsubsidized work. For welfare recipients, NSW significantly increased unsubsidized employment and reduced welfare receipt.[56]

Also relevant is the Aid to Families with Dependent Children (AFDC) Homemaker-Home Health Aide Demonstration (HHAD), which is described as "comparable in duration and intensity to the National Supported Work Demonstration."[57] The program provided four to eight weeks of formal classroom training in being a home health aide and up to 12 months of transitional jobs to individuals who had been on welfare for at least three consecutive months in seven states (Arkansas, Kentucky, New Jersey, New York, Ohio, South Carolina, and Texas). Like TWC, participation was high: about 85 percent of the program group received training and 72 percent participated in transitional jobs. In the year after individuals were generally no longer in training or working in transitional jobs, the programs increased employment, hours worked, and earnings.[58] The programs also reduced monthly AFDC and food stamp benefits.[59] In the second post-demonstration year, effects on employment-related outcomes were even larger, but effects on welfare benefits declined.

One problem with using HHAD as a prior for TWC is the program's use of classroom training: 85 percent of the HHAD program groups received some training, but only 33.9 percent of the TWC group participated in any education or training activity.[60] Because classroom training has been found to increase the earnings of welfare recipients, the longer-term effects of these programs may be a result of training rather than transitional jobs.[61]

Also potentially relevant to TWC are two random assignment studies of voluntary on-the-job training (OJT) for welfare recipients. The Maine Training Opportunities in the Private Sector (TOPS) program was a small-scale, voluntary program for welfare recipients who had

[55] Bloom (2010).
[56] Manpower Demonstration Research Corporation Board of Directors (1980).
[57] Bell and Orr (1994).
[58] Bell and Orr (1994); Enns, Bell, and Flanagain (1987).
[59] Enns, Bell, and Flanagain (1987).
[60] Bell and Orr (1994); Bloom et al. (2009).
[61] Greenberg, Michalopoulos, and Robins (2003).

been on the rolls for at least six months and were not working.[62] It contained three program elements: two to five weeks of pre-vocational training that stressed job-seeking and job-holding skills (which sounds similar to the training provided in TWC); up to 12 weeks of unpaid, part-time work experience in the public sector or nonprofit sector; and on-the-job training — primarily in the private sector — that paid employers half of the individual's salary for up to six months. As in TWC, individuals received a paycheck when they were in subsidized jobs and their welfare grants were reduced as they would have been for any employed individual. Participation was high: 88.6 percent participated in pre-vocational training, 67.7 percent in unpaid work experience, and 28.6 percent in OJT.

In the New Jersey Grant Diversion Project, private-sector employers hired AFDC recipients, mostly single mothers, for up to six months with the understanding that those who performed satisfactorily would be kept on as regular full-time employees.[63] The state paid half the person's wages during the six-month OJT period. About 40 percent of program group members worked in an OJT position, which is somewhat smaller than in TWC.

Another set of studies that are similar in some respects to TWC are the state Community Work Experience Programs evaluated by MDRC in West Virginia, Chicago,[64] and San Diego. Each program required welfare recipients to work in order to continue receiving benefits. Although such work was a requirement, participation rates in these jobs were quite low, ranging from 7 percent in Chicago to 24 percent in West Virginia. These programs were, therefore, not considered in developing priors for TWC.[65]

Table 8 shows results from these sets of studies. Because the transitional jobs programs were not included in the reported employment measures in all of the programs, the four sets of estimated effects on employment vary substantially across the studies. In particular, TOPS and HHAD show reductions in employment in the first year. In addition, the HHAD reports provide information only on annual employment. By contrast, the New Jersey OJT study increased total employment by more than 10 percentage points initially, while NSW increased employment by more than 75 percentage points in the first quarter. This may be another indication that NSW is not a good match for TWC; the TWC evaluation found increases in total employment of about only 25 percentage points while individuals were eligible for transitional jobs.

[62] Auspos, Cave, and Long (1988).
[63] Freedman, Bryant, and Cave (1988).
[64] The study took place throughout Cook County, of which Chicago is the county seat.
[65] Friedlander (1986, 1987); Goldman (1985).

Enhanced Services for the Hard-to-Employ Demonstration

Table 8

Impacts on Employment and Welfare Receipt, Subsidized Employment for Welfare Recipients

Bayesian Reanalysis

	Maine TOPS		New Jersey OJT		NSW		AFDC HHAD Demo	
	Impact	Standard	Impact	Standard	Impact	Standard	Impact	Standard
Outcome (%)	Estimate	Error	Estimate	Error	Estimate	Error	Estimate	Error
Employment								
Quarter 1	-8.3	4.2	15.3	2.5	63.0	2.1	-12.4	1.1
Quarter 2	-11.3	4.6	13.1	2.5	54.2	2.3		
Quarter 3	3.5	4.9	4.5	2.5	46.0	2.4		
Quarter 4	8.3	5.0	2.5	2.5	26.7	2.6		
Quarter 5	8.6	5.0	5.3	3.2	5.3	2.6	3.0	1.6
Quarter 6	7.2	5.0	-0.2	3.1	8.1	3.9		
Quarter 7	5.4	5.0	-1.7	3.1	7.4	3.9		
Quarter 8	6.9	5.0			7.1	3.9		
Quarter 9	7.1	5.0					10.6	1.7
Quarter 10	11.1	5.0						
Quarter 11	1.1	5.0						
Welfare receipt								
Quarter 1	-0.7	0.9	1.4	1.0				
Quarter 2	-0.4	1.5	-2.2	1.4				
Quarter 3	0.0	3.0	-5.1	2.1				
Quarter 4	-3.3	3.7	-5.6	2.4				
Quarter 5	0.2	4.2	-5.9	2.5				
Quarter 6	2.4	4.5	-6.4	3.2				
Quarter 7	1.3	4.7	-3.8	3.2				
Quarter 8	1.3	4.9	-1.9	3.2				
Quarter 9	2.5	4.9						
Quarter 10	6.4	5.0						
Quarter 11	6.0	5.0						
Sample size	444		1,604		1,315		3,378	

SOURCES: Auspos, Cave, and Long (1988) for Maine Training Opportunities in the Private Sector (TOPS) estimates; Freedman, Bryant, and Cave (1988) for New Jersey On-the-Job-Training (OJT) estimates; MDRC Board of Directors (1980) for National Supported Work (NSW) Demonstration estimates; and Enns, Bell, and Flanagain (1987) for AFDC Homemaker-Home Health Aide Demonstration (HHAD) estimates.

NOTES: Only 7 quarters of employment results and 8 quarters of welfare receipt results were published for the New Jersey OJT study. Only 9 quarters of employment results were published for NSW, but because transitional jobs in NSW lasted about three months more on average than in TWC, the first quarter of results from NSW were not used in forming the prior distribution. Published impacts on welfare receipt were not available for NSW and AFDC HHAD. Published results for AFDC HHAD showed employment impacts in 9-month periods. Results for months 1-9 are shown in Quarter 1 in the table, results for months 10-18 are shown in Quarter 5, and results for months 19-27 are shown in Quarter 9.

In the second and third years, after individuals were no longer eligible for transitional jobs, all of the programs other than New Jersey showed some evidence of ongoing employment gains. In Maine, these estimated effects were as large as 11 percentage points and generally were in the 7 to 9 percentage point range, although increases in earnings were significantly different from zero only in Quarters 5 and 10. Impacts in the larger NSW and HHAD studies were consistent with results from Maine, ranging between 5.3 and 8.1 percentage points from Quarter 6 on in NSW and as high as 10.6 percentage points in the AFDC study in Year 3.

The studies vary substantially in their longer-term effects on welfare receipt, with no significant impacts in Maine and modest initial reductions in New Jersey that were close to zero by Quarter 8. Results on welfare receipt were not reported in the HHAD reports, which documented benefit amounts but not receipt rates.

Results of the Bayesian Analysis for TWC

For TWC, the Bayesian analysis addresses two questions: (1) Do transitional jobs lead to longer-term effects on unsubsidized employment? (2) Do they lead to longer-term effects on welfare receipt? Because there is some question about which studies are directly relevant to understanding the effects of TWC, two sets of priors using the earlier studies are investigated. One set includes all four of the studies described earlier: Maine, New Jersey, NSW, and HHAD. Because HHAD included more substantial vocational training than TWC and because NSW had substantially greater use of transitional jobs than TWC, a second set of priors is based on results from Maine and New Jersey only.

The Bayesian results for employment and welfare receipt are presented in Table 9. Because the different studies reported employment differently while individuals were eligible for transitional jobs, results are presented beginning in Quarter 5.

The results tell a fairly clear story of declining impacts and suggest that TWC probably increased employment and reduced welfare receipt by a fair amount at the beginning of the second year, but had little effect after that. In Quarter 5, for example, the mean estimated effect on employment is a 6.4 percentage point increase and on welfare receipt is a 5.5 percentage point reduction. Moreover, there is a 98.9 percent probability that TWC increased employment and a 99.1 percent probability that it reduced welfare receipt during this quarter. Through the second year, the probability that TWC had a positive effect on employment declined to 56.2 percent, but by the beginning of the third year, the evidence suggests that TWC actually might have reduced employment and increased welfare receipt.

The results vary substantially depending on which priors are used. Results are much more positive when priors from the other four studies are included. The results generally suggest that TWC increased employment and reduced welfare receipt through Quarter 8, with

Enhanced Services for the Hard-to-Employ Demonstration

Table 9

Estimated Effects on Employment and Welfare Receipt, Subsidized Employment for Welfare Recipients: Summary of Bayesian Posterior Distributions

Bayesian Reanalysis

	Uninformative Priors			Priors Based on 4 Studies[a]			Priors Based on ME and NJ[b]		
Outcome (%)	Mean Estimated Effect	Standard Deviation	Probability of Positive Impact	Mean Estimated Effect	Standard Deviation	Probability of Positive Impact	Mean Estimated Effect	Standard Deviation	Probability of Positive Impact
Employment									
Quarter 5	6.4	2.8	98.9	4.5	1.1	100.0	6.3	1.9	99.9
Quarter 6	1.8	2.8	74.2	3.1	1.2	99.5	1.9	1.9	83.4
Quarter 7	1.3	2.8	67.4	2.6	1.2	98.6	0.8	1.9	65.2
Quarter 8	0.4	2.8	56.2	3.2	1.3	99.4	2.0	2.5	79.2
Quarter 9	-5.7	2.8	2.2	6.5	1.4	100.0	-2.6	2.5	14.6
Welfare receipt[c]									
Quarter 5	-5.5	2.4	0.9	-4.9	1.6	0.1	-4.9	1.6	0.1
Quarter 6	-6.9	2.5	0.3	-5.3	1.8	0.2	-5.3	1.8	0.2
Quarter 7	-1.9	2.6	23.0	-2.0	1.8	13.3	-2.0	1.8	13.3
Quarter 8	-2.4	2.6	17.9	-1.7	1.9	18.3	-1.7	1.9	18.3
Quarter 9	0.9	2.7	62.8	1.2	2.4	70.2	1.2	2.4	70.2
Sample size (total = 1,942)									

SOURCES: Jacobs and Bloom (2011) and Bayesian results from author calculations.

NOTES: [a]Maine Training Opportunities in the Private Sector (TOPS) study, New Jersey On-the-Job-Training (OJT) project, National Supported Work (NSW) Demonstration, and AFDC Homemaker-Home Health Aide Demonstration (HHAD).

[b]Maine TOPS study and New Jersey OJT project.

[c]Estimated impacts on welfare receipt were not available from published reports on NSW and the AFDC HHAD studies. Therefore, welfare receipt estimates are the same using priors based on the four studies and using priors based on just ME and NJ.

mean effects on employment ranging from 2.6 to 6.5 percentage points and mean effects on welfare receipt ranging from 1.7 to 4.9 percentage points. The probability that TWC increased employment is nearly 100 percent in each quarter, and the probability that TWC reduced welfare receipt is more than 80 percent in four of the five quarters shown in Table 9.

The last set of results, which use only the Maine and New Jersey studies to form a prior, tell a less positive story and look quite similar to the results with an uninformative prior.[66] The mean estimated effect on employment declines over time from 6.3 percentage points in Quarter 5 to −2.6 percentage points in Quarter 9, and the probability that TWC increased employment declines from 99.9 percent to 14.6 percent. Likewise, the mean estimated effect on welfare receipt declines from about 5 percentage points in Quarters 5 and 6 to an increase of 1.2 percentage points in Quarter 9, and the probability that TWC reduced welfare receipt declines from nearly 100 percent in Quarters 5 and 6 to 29.8 percent in Quarter 9.

The similarity of these results to the results using the uninformative prior indicates that results from TWC confirmed what had been found in the Maine and New Jersey studies. Transitional jobs may have longer-term effects if NSW and HHAD are comparable to TWC. However, the presence of vocational training in those two studies and the absence of such training for the most part in the other three studies may imply that the more positive long-term effects from NSW and HHAD are a result of that training, either alone or in combination with transitional jobs.

Transitional Jobs: Discussion

The Bayesian reanalysis for CEO suggests that the transitional jobs program for ex-prisoners probably caused a small increase in unsubsidized employment and a small decrease in recidivism. Results for the recently released subgroup are much smaller than those in the published CEO results, which used a classical analysis, especially regarding recidivism. This is because estimates from the Hard-to-Employ evaluation of CEO were larger than in TJRD or NSW for ex-prisoners. The smaller estimates from the other studies casts some skepticism on the size of CEO's effects unless there is a persuasive explanation for how CEO differs from the other studies or CEO's results can be replicated.

With regard to the TWC transitional jobs program for welfare recipients, the analysis shows how sensitive the results can be to the manner in which Bayesian priors are chosen. Results from NSW and HHAD were so positive that, when combined with new results from TWC, they suggest that transitional jobs are very likely to have modest effects on employment

[66] Starting in Quarter 5, estimated effects on welfare receipt using this set of priors are the same as results using the set of priors from the four studies, because impacts on welfare receipt were not available in NSW or HHAD.

and welfare receipt in the medium term. If those programs are not considered comparable to TWC, however, perhaps because they included more intensive training components, the Bayesian analysis confirms the TWC evaluation findings that the program had little effect on unsubsidized employment and welfare receipt in the medium term.

Conclusion

This paper has presented a reanalysis of results from three sites of the Hard-to-Employ evaluation using a Bayesian perspective. The paper had two goals: (1) to illustrate how Bayesian analyses might be used in social policy evaluations, and (2) to assess findings from the three sites in light of results from similar studies.

For WtW, the reanalysis suggested that the program's effects were likely to be small but positive. This is a more positive interpretation than indicated by the published classical analysis. For example, the reanalysis suggests that the Working toward Wellness program probably reduced depression slightly, while the original analysis indicated that the impact on depression severity was not statistically significant. The Bayesian reanalysis provided greater confidence in small impacts for two reasons. First, results were generally consistent with results from other telephonic care management studies. Formally incorporating evidence from those studies using a Bayesian framework reduced uncertainty and provided greater confidence that small impacts were likely to represent true program effects. Second, the Bayesian reanalysis resulted in a reinterpretation of results that focused less on whether impacts were statistically significant but instead provided more information on the likely distribution of impacts.

For CEO, the reanalysis suggested that the program's effects were likely to be very small and the reductions in recidivism may have been much smaller than suggested by the published CEO results. This difference occurs because transitional jobs programs in the Transitional Jobs Reentry Demonstration and the National Supported Work Demonstration showed little effect on recidivism in the second year. The Bayesian reanalysis, therefore, provides a reason to be skeptical about CEO's effects on recidivism, and the ongoing replication of CEO will provide important information on the program's likely true effects.

For the transitional jobs program in Philadelphia, it was more difficult to settle on good comparison studies. This is in part because prior evaluations of subsidized employment for welfare recipients occurred decades ago, but also because earlier studies differed in important ways — for example, by including more intensive training opportunities than were provided by the Philadelphia program. Because of this uncertainty, the Bayesian reanalysis was less conclusive. The transitional jobs approach may or may not have produced modest employment gains and reductions in welfare receipt in the medium term.

This paper suggests three ways in which social policy evaluations might be altered. First, evaluations may want to put more effort into placing their results into the context of earlier findings. In many reports and papers, this is done in a literature review section near the beginning of the document. It may also be illuminating to include a later section that explicitly compares results across studies. This may be especially important for key policy findings or for surprising or questionable results.

Second, evaluations may want to place less emphasis on statistical significance tests and more emphasis on the implied distribution of estimated effects. While most academic journals require some information about statistical significance, this may not be the most useful standard for making policy decisions. A policymaker may benefit more from knowing there is an 85 percent chance that a program's effect is positive than from knowing the estimate is not significantly different from zero at conventional significance levels. This can be done within the standard statistical framework by discussing confidence intervals for key impact estimates.

Finally, a Bayesian framework may allow for smaller evaluations in some cases. Where evidence exists to form a strong prior, for example, Bayesian updating can be used to obtain precise statistical inferences. This may be the case in deciding whether to expand a policy that has been evaluated in one set of sites or for one target population to a different location or group. A small study interpreted from a Bayesian perspective may also be sufficient if policymakers need to know only that the preponderance of evidence is positive, but do not need to see statistically significant differences. One example is in making sure there is little evidence of reduced effectiveness when deciding to use a less expensive form of an intervention.

While not often used in the social sciences, this paper has illustrated how a Bayesian statistical analysis adds perspective to traditional evaluation methods by placing evaluation data in the context of existing research evidence. In this way, Bayesian analysis has the potential to improve social sciences evaluation and should be considered, particularly when there is a large body of relevant studies to provide that context or when a new study is smaller than needed to detect statistically significant impacts. The Bayesian method for drawing inferences likewise has particular relevance to studies where it is not enough to know whether an impact is likely to be zero.

References

Adelmann, Pamela K. 2003. "Mental and Substance Use Disorders Among Medicaid Recipients: Prevalence Estimates from Two National Surveys." *Administration and Policy in Mental Health* 31, 2: 111-129.

Auspos, Patricia, George Cave, and David Long with Karla Hanson, Emma Caspar, Daniel Friedlander, and Barbara Goldman. 1988. *Maine: Final Report on the Training Opportunities in the Private Sector Program*. New York: MDRC.

Bell, Stephen H., and Larry L. Orr. 1994. "Is Subsidized Employment Cost Effective for Welfare Recipients? Experimental Evidence from Seven State Demonstrations." *Journal of Human Resources* 29, 1: 42-61.

Belsher, Gayle, and Charles G. Costello. 1988. "Relapse After Recovery from Unipolar Depression: A Critical Review." *Psychological Bulletin* 104, 1: 84-96.

Berk, Richard A., Alec Campbell, Ruth Klap, and Bruce Western. 1992. "A Bayesian Analysis of the Colorado Springs Spouse Abuse Experiment." *The Journal of Criminal Law and Criminology* 83, 1: 170-200.

Berry, Donald A. 1996. *Statistics: A Bayesian Perspective.* Belmont, CA: Duxbury Press.

Bloom, Dan. 2010. *Transitional Jobs: Background, Program Models, and Evaluation Evidence*. New York: MDRC.

Bloom, Dan, Cindy Redcross, Janine Zweig, and Gilda Azurdia. 2007. *Transitional Jobs for Ex-Prisoners: Early Impacts from a Random Assignment Evaluation of the Center for Employment Opportunities (CEO) Prisoner Reentry Program*. New York: MDRC.

Bloom, Dan, Sarah Rich, Cindy Redcross, Erin Jacobs, Jennifer Yahner, and Nancy Pindus. 2009. *Alternative Welfare-to-Work Strategies for the Hard-to-Employ: Testing Transitional Jobs and Pre-Employment Services in Philadelphia*. New York: MDRC.

Bolstad, William M. 2007. *Introduction to Bayesian Statistics.* Hoboken, NJ: John Wiley & Sons, Inc.

Breslow, Norman. 1990. "Biostatistics and Bayes." *Statistical Science* 5, 3: 269-284.

Bushway, Shawn, and Peter Reuter. 2002. "Labor Markets and Crime." In J. Q. Wilson and J. Petersilia (eds.), *Crime: Public Policies and Crime Control*. Oakland, CA: Institute for Contemporary Studies Press.

Danziger, Sandra, Mary Corcoran, Sheldon Danziger, Colleen Heflin, Ariel Kalil, Judith Levine, Daniel Rosen, Kristin Seefeldt, Kristine Siefert, and Richard Tolman. 2002. "Barriers to the Employment of Welfare Recipients." PSC Research Report No. 02-508. Ann Arbor: University of Michigan, Population Studies Center at the Institute for Social Research.

Enns, John H., Stephen H. Bell, and Kathlen L. Flanagain. 1987. *AFDC Homemaker-Home Health Aide Demonstrations: Trainee Employment and Earnings*. Bethesda, MD: Abt Associates Inc.

Freedman, Stephen, Jan Bryant, and George Cave. 1988. *New Jersey: Final Report on the Grant Diversion Project*. New York: MDRC.

Friedlander, Daniel. 1987. *Final Report on Job Search and Work Experience in Cook County*. New York: MDRC.

Friedlander, Daniel. 1986. *West Virginia: Final Report on the Community Work Experience Demonstrations*. New York: MDRC.

Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 2004. *Bayesian Data Analysis.* Boca Raton, FL: Chapman & Hall.

Gennetian, Lisa A. 2003. *The Long-Term Effects of the Minnesota Family Investment Program on Marriage and Divorce Among Two-Parent Families*. New York: MDRC.

Gigerenzer, Gird. 2002. *Calculated Risks: How to Know When Numbers Deceive You.* New York: Simon & Schuster.

Goldman, Barbara. 1985. *Findings from the San Diego Job Search and Work Experience Demonstration*. New York: MDRC.

Greenberg, David H., Charles Michalopoulos, and Philip K. Robins. 2003. "A Meta-Analysis of Government Sponsored Training Programs." *Industrial and Labor Relations Review* 50, 1: 31-53.

Jacobs, Erin. 2012. *Returning to Work After Prison: Final Results from the Transitional Jobs Reentry Demonstration.* New York: MDRC.

Jacobs, Erin, and Dan Bloom. 2011. *Alternative Employment Strategies for Hard-to-Employ TANF Recipients: Final Results from a Test of Transitional Jobs and Preemployment Services in Philadelphia.* New York: MDRC.

Kim, Sue, Allen LeBlanc, and Charles Michalopoulos. 2009. *Working toward Wellness: Early Results from a Telephone Care Management Program for Medicaid Recipients with Depression.* New York: MDRC.

Ioannidis, John P. A. 2005. "Why Most Published Research Findings Are False." *PLoSMedicine* 2, 8: 696-701.

Laub, John H., and Robert J. Sampson. 2001. "Understanding Desistance from Crime." Pages 1-69 in M. Tonry (ed.), *Crime and Justice: A Review of Research.*Volume 28. Chicago: University of Chicago Press.

Lilford, Richard, and David Braumholtz. 1996. "The Statistical Bias in Public Policy: A Paradigm Shift Is Overdue." *British Medical Journal* 313: 603-607.

Manpower Demonstration Research Corporation Board of Directors. 1980. *Summary and Findings of the National Supported Work Demonstration.* Cambridge, Massachusetts: Ballinger Publishing Company.

McGrayne, Sharon Bertsch. 2011. *The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries.* New Haven: Yale University Press.

Miller, Cynthia, Virginia Knox, Lisa A. Gennetian, Martey Dodoo, Jo Anna Hunter, and Cindy Redcross. 2000. *Reforming Welfare and Rewarding Work: Final Report on the Minnesota Family Investment Program, Volume 1, Effects on Adults*. New York: MDRC.

Miranda, Jeanne, Joyce Y. Chung, Bonnie L. Green, Janice Krupnick, Juned Siddique, Dennis A. Revicki, and Tom Belin. 2003. "Treating Depression in Predominantly Low-Income Young Minority Women: A Randomized Controlled Trial." *Journal of the American Medical Association* 290, 1: 57-65.

Miranda, Jeanne, Bonnie L. Green, Janice L. Krupnick, Joyce Chung, Juned Siddique, Tom Belin, and Dennis Revicki. 2006. "One-Year Outcomes of a Randomized Clinical Trial Treating Depression in Low-Income Minority Women." *Journal of Consulting and Clinical Psychology* 74, 1: 99-111.

Miranda, Jeanne, Michael Schoenbaum, Cathy Sherbourne, Naihua Duan, and Kenneth Wells. 2004. "Effects of Primary Care Depression Treatment on Minority Patients' Clinical Status and Employment." *Archives of General Psychiatry* 61, 8: 827-834.

Redcross, Cindy, Dan Bloom, Gilda Azurdia, Janine Zweig, and Nancy Pindus. 2009. *Transitional Jobs for Ex-Prisoners: Implementation, Two-Year Impacts, and Costs of the Center for Employment Opportunities (CEO) Prisoner Reentry Program*. New York: MDRC.

Redcross, Cindy, Dan Bloom, Erin Jacobs, Michelle Manno, Sara Muller-Ravett, Kristin Seefeldt, Jennifer Yahner, Jr. Alfred A. Young, and Janine Zweig. 2010. *Work After Prison: One-Year Findings from the Transitional Jobs Reentry Demonstration*. New York: MDRC.

Redcross, Cindy, Megan Millenky, Tim Rudd, and Valerie Levshin. 2012. *More Than a Job: Final Results from the Evaluation of the Center for Employment Opportunities (CEO) Transitional Jobs Program.* New York: MDRC.

Sherman, Lawrence W. and Richard A. Berk. April 1984. "The Minneapolis Domestic Violence Experiment." *Police Foundation Reports*. Washington: Police Foundation.

Simon, Gregory E., Evette J. Ludman, Steve Tutty, Belinda Operskalski, and Michael VonKorff. 2004. "Telephone Psychotherapy and Telephone Care Management for Primary Care Patients Starting Antidepressant Treatment: A Randomized Controlled Trial." *Journal of the American Medical Association* 292, 8: 935-942.

Simon, Gregory E., Michael VonKorff, Carolyn Rutter, and Edward Wagner. 2000. “Randomised Trial of Monitoring, Feedback, and Management of Care by Telephone to Improve Treatment of Depression in Primary Care.” *BMJ* 320, 7234: 550-554.

Spiegelhalter, David J., Laurence S. Freedman, and Mahesh K.B. Parmar. 1994. “Bayesian Approaches to Randomized Trials.” *Journal of the Royal Statistical Society* 157, 3: 357-416.

Spiegelhalter, David J., Jonathan P. Myles, David R. Jones, and Keith R. Abrams. 2000. “Bayesian Methods in Health Technology Assessment: A Review.” *Health Technology Assessment* 4, 38.

Vail, Andy, Janet Hornbuckle, David J. Spiegelhalter, and Jim G. Thornton. 2001. “Prospective Application of Bayesian Monitoring and Analysis in an ‘Open’ Randomized Clinical Trial.” *Statistics in Medicine* 20: 3777-3787.

Wang, Philip S., Gregory E. Simon, Jerry Avorn, Francisca Azocar, Evette J. Ludman, Joyce McCulloch, Maria Z. Petukhova, and Ronald C. Kessler. 2007. “Telephone Screening, Outreach, and Care Management for Depressed Workers and Impact on Clinical and Work Productivity Outcomes: A Randomized Controlled Trial.” *Journal of the American Medical Association* 298, 12: 1401-1411.

Wang, Philip S., Arne L. Beck, Pat Berglund, David K. McKenas, Nicolaas Pronk, Gregory E. Simon, and Ronald C. Kessler. 2004. “Effects of Major Depression on Moment-in-Time Work Performance.” *American Journal of Psychiatry* 16, 10: 1885-1891.

Wells, Kenneth B., Cathy Sherbourne, Michael Schoenbaum, Naihua Duan, Lisa Meredith, Jürgen Unutzer, Jeanne Miranda, Maureen F. Carney, and Lisa V. Rubenstein. 2000. “Impact of Disseminating Quality Improvement Programs for Depression in Managed Primary Care: A Randomized Controlled Trial.” *Journal of the American Medical Association* 283, 2: 212-220.

Wilson, David B., Catherine A. Gallagher, and Doris L. MacKenzie. 2000. “A Meta-Analysis of Corrections-Based Education, Vocation, and Work Programs for Adult Offenders.” *Journal of Research in Crime and Delinquency* 37, 4: 347-368.

MDRC Publications on the Enhanced Services for the Hard-to-Employ Demonstration and Evaluation Project

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
2012. David Butler, Julianna Alson, Dan Bloom, Victoria Deitch, Aaron Hill, JoAnn Hsueh, Erin Jacobs, Sue Kim, Reanin McRoberts, and Cindy Redcross.

Enhanced Early Head Start with Employment Services: 42-Month Impacts from the Kansas and Missouri Sites of the Enhanced Services for the Hard-to-Employ Demonstration and Evaluation Project
2012. Joann Hsueh and Mary E. Farrell.

Investigating Depression Severity in the Working toward Wellness Study
2012. Sue Kim and Charles Michalopoulos.

More Than a Job: Final Results from the Evaluation of the Center for Employment Opportunities (CEO) Transitional Jobs Program
2012. Cindy Redcross, Megan Millenky, Timothy Rudd, and Valerie Levshin.

Alternative Employment Strategies for Hard-to-Employ TANF Recipients: Final Results from a Test of Transitional Jobs and Preemployment Services in Philadelphia
2011. Erin Jacobs and Dan Bloom.

Working toward Wellness: Telephone Care Management for Medicaid Recipients with Depression, Thirty-Six Months After Depression
2011. Sue Kim, Allen LeBlanc, Pamela Morris, Greg Simon, and Johanna Walter.

A Two-Generational Child-Focused Program Enhanced with Employment Services: Eighteen-Month Impacts from the Kansas and Missouri Sites of the Enhanced Services for the Hard-to-Employ Demonstration and Evaluation Project
2011. JoAnn Hsueh, Erin Jacobs, and Mary Farrell.

Working toward Wellness: Telephone Care Management for Medicaid Recipients with Depression, Eighteen Months After Random Assignment
2010. Sue Kim, Allen LeBlanc, Pamela Morris, Greg Simon, and Johanna Walter.

Recidivism Effects of the Center for Employment Opportunities (CEO) Program Vary by Former Prisoners' Risk of Reoffending
2010. Janine Zweig, Jennifer Yahner, and Cindy Redcross.

Transitional Jobs: Background, Program Models, and Evaluation Evidence
2010. Dan Bloom.

Alternative Welfare-to-Work Strategies for the Hard-to-Employ: Testing Transitional Jobs and Pre-Employment Services in Philadelphia
2009. Dan Bloom, Sarah Rich, Cindy Redcross, Erin Jacobs, Jennifer Yahner, and Nancy Pindus.

Working toward Wellness: Early Results from a Telephone Care Management Program for Medicaid Recipients with Depression
2009. Sue Kim, Allen LeBlanc, and Charles Michalopoulos.

Transitional Jobs for Ex-Prisoners: Implementation, Two-Year Impacts, and Costs of the Center for Employment Opportunities (CEO) Prisoner Reentry Program
2009. Cindy Redcross, Dan Bloom, Gilda Azurdia, Janine Zweig, and Nancy Pindus.

"Transitional Jobs for Ex-Prisoners: Early Impacts from a Random Assignment Evaluation of the Center for Employment Opportunities (CEO) Prisoner Reentry Program." Working Paper.
2007. Dan Bloom, Cindy Redcross, Janine Zweig, and Gilda Azurdia.

Four Strategies to Overcome Barriers to Employment: An Introduction to the Enhanced Services for the Hard-to-Employ Demonstration and Evaluation Project
2007. Dan Bloom, Cindy Redcross, JoAnn Hsueh, Sarah Rich, and Vanessa Martin.

The Power of Work: The Center for Employment Opportunities Comprehensive Prisoner Reentry Program
2006. The Center for Employment Opportunities and MDRC.