Evaluation of
Employment Coaching for TANF and Related Populations

Using Bayesian Methods to Conduct Subgroup Analysis in Evaluations of Employment Programs

OPRE Report Number 2024-027 • February 2024

Using Bayesian Methods to Conduct Subgroup Analysis in Evaluations of Employment Programs

OPRE Report Number 2024-027 • February 2024

Tim Kautz, Christina Kent, and Dan Thal

Submitted to:
Office of Planning, Research, and Evaluation
Administration for Children and Families
U.S. Department of Health and Human Services
330 C Street, SW
Washington, DC 20201
Project Officers: Hilary Bruck, Lauren Deutsch Stanton, Sarita Barton, and Elizabeth Karberg

Contract/Task Number: 140D0421F0748

Mathematica Reference Number: 51304

Submitted by:
Mathematica
1100 1st Street, NE
12th Floor
Washington, DC 20002-4221
Telephone: (202) 484-9220
Facsimile: (202) 863-1763

Acknowledgements

We appreciate the staff at the Office of Planning, Research, and Evaluation (OPRE) who provided support, feedback, and advice for this study, including Hilary Bruck, Lauren Deutsch Stanton, Sarita Barton, and Elizabeth Karberg. We also benefited from helpful comments from John Deke, Rebecca Connelly Kersting, Mariel Finucane, Sheena McConnell, Quinn Moore, and Rob Wood. Evan Morier and Charles Tilley provided excellent programming support. Bridget Gutierrez provided editorial assistance, and Asa Wild, Laura Sarnoski, and Yvonne Marki provided production support.

Contents

Tables

Overview

Many Temporary Assistance for Needy Families (TANF) recipients and other individuals with low incomes seek employment and training programs to help them find jobs or improve their earnings, which could, in turn, allow them to better support their families. However, these programs do not necessarily benefit all participants equally. Program evaluations that include subgroup analysis can inform how employment programs provide their services and help them improve equity by identifying who needs more tailored services. This report details new and promising approaches to subgroup analysis for evaluators of employment programs. It discusses how two Bayesian methods—a Bayesian hierarchical linear model and a Bayesian causal forest—can potentially address limitations of standard subgroup analysis. The report uses data from four experimental evaluations of employment programs in the Evaluation of Employment Coaching for TANF and Related Populations, a project sponsored by the Office of Planning, Research, and Evaluation in the Administration for Children and Families, U.S. Department of Health and Human Services. The results suggest that Bayesian methods can complement traditional methods of conducting subgroup analyses in impact evaluations.

PRIMARY RESEARCH QUESTIONS

The report addresses three research questions:

1. How can Bayesian methods be used to reinterpret subgroup impact estimates for a single evaluation of an employment program?
2. How can Bayesian methods be used to reinterpret subgroup impact estimates for multiple evaluations of employment programs?
3. To what extent can Bayesian methods be used in an evaluation of an employment program to identify subgroups that were not previously specified?

PURPOSE

The purpose of this study is to inform evaluators of employment programs about new and promising approaches to subgroup analysis by illustrating how Bayesian methods for subgroup analysis can complement traditional methods. Drawing on real-world data, the study provides evidence on (1) how Bayesian methods can be used to reinterpret subgroup impact estimates for a single evaluation; (2) how Bayesian methods can be used to reinterpret subgroup impact estimates for multiple evaluations; and (3) the extent to which Bayesian methods can be used to identify subgroups that were not previously specified by the evaluator conducting the analysis. The study aims to identify methods that can draw more nuanced insights from subgroup analyses in the context of evaluations. Such insights can help practitioners and policymakers better serve and design programs for specific groups.

KEY FINDINGS AND HIGHLIGHTS

The report found that:

- Compared to null hypothesis testing, the Bayesian hierarchical linear model approach can (1) suggest more nuanced conclusions about the differences between subgroups and (2) lead to estimates that are less sensitive to small deviations in the data.
- When applying the Bayesian hierarchical linear model approach to multiple evaluations of employment programs, the impact estimates from separate programs influence each other, highlighting how a Bayesian hierarchical linear model draws on information across programs.
- The Bayesian causal forest approach can potentially identify subgroups that had not been pre-specified by the evaluator conducting the analyses in evaluations with large sample sizes.

METHODS

The report includes:

- A traditional subgroup impact analysis based on linear regressions and null hypothesis testing
- A Bayesian hierarchical linear model approach to reinterpreting traditional impact estimates
- A Bayesian causal forest approach to identifying subgroups that were not previously specified by the evaluator conducting the analyses

Executive Summary

Many Temporary Assistance for Needy Families (TANF) recipients and other individuals with low incomes seek employment and training programs to help them find jobs or improve their earnings, which could, in turn, allow them to better support their families. However, these programs do not necessarily benefit all participants equally. In evaluations of employment programs, subgroup analysis informs whether program impacts on outcomes differ based on group characteristics (such as gender or age), suggesting whether participants with certain characteristics benefit from program participation more than others. The findings from a subgroup analysis can inform how employment programs provide their services and help them improve equity.

The standard approach to subgroup analysis involves calculating separate impact estimates for groups of participants and conducting null hypothesis tests to determine whether the difference in impact estimates between the groups is statistically significant. These standard subgroup analyses are subject to three limitations. First, the results from null hypothesis tests can be hard to interpret for practitioners and policymakers (Greenland et al. 2016; Wasserstein and Lazar 2016; Gigerenzer 2018). In addition, they provide limited information on whether differences in impacts across subgroups are meaningful (Goodman 2016). Second, small changes to the analysis or data may affect whether the results are statistically significant and thus change the conclusions drawn from the data (Gelman and Stern 2006; Gelman 2013). Third, impact studies are often designed to detect effects when using the overall study sample and not to detect effects for subgroups that make up a part of the sample. Therefore, if policymakers and practitioners are interested in a subgroup that is a small subset of the overall sample, the analyses may be underpowered.

To guide evaluators of employment programs, this report discusses how Bayesian methods can potentially address limitations of standard subgroup analysis. It uses real-world data to illustrate how two Bayesian methods—a Bayesian hierarchical linear model and a Bayesian causal forest —can complement the typical approach to subgroup analysis in evaluations of employment programs. To explore these methods, the report uses data from four experimental evaluations of employment programs in the Evaluation of Employment Coaching for TANF and Related Populations (Evaluation of Employment Coaching), a project sponsored by the Office of Planning, Research, and Evaluation in the Administration for Children and Families, U.S. Department of Health and Human Services.

OVERVIEW OF THE BAYESIAN APPROACHES

This report discusses two Bayesian approaches for analyzing subgroups in employment program evaluations:

1. A Bayesian hierarchical linear model allows evaluators to reinterpret the original subgroup estimates in terms of probabilities rather than statistical significance, thereby providing more nuanced conclusions. For example, for an impact estimate that was not statistically significant, a Bayesian hierarchical linear model could still indicate whether the impact was likely to be positive. This finding allows evaluators to make statements such as, "There was an 80 percent chance that the impact on earnings for males exceeded $0."
2. A Bayesian causal forest provides a flexible way to discover meaningful subgroups that had not been previously considered or specified by the evaluator conducting the analysis (Hahn et al. 2020). The Bayesian causal forest approach could potentially identify subgroups for which impacts differed based on a complex combination of variables that evaluators may not consider. For example, it could reveal that an employment program conferred especially large benefits to women younger than age 27 who do not live with a child.

PERFORMANCE OF THE BAYESIAN APPROACHES

The Bayesian hierarchical linear model approach could complement standard null hypothesis testing for a single evaluation by accounting for multiple comparisons, providing more nuanced conclusions, and guarding against small changes in the data or modeling decisions.

This report illustrates several ways that evaluators could use Bayesian approaches to explore subgroup impacts of employment programs (Table 1). Our results show how the Bayesian hierarchical linear model approach could complement standard null hypothesis testing for a single evaluation by accounting for multiple comparisons, providing more nuanced conclusions, and guarding against small changes in the data or modeling decisions. The resulting probability statements allow evaluators to provide some information even when results do not meet standard levels of statistical significance.

In addition, our analysis shows how a Bayesian hierarchical linear model could be used across multiple evaluations of employment programs or evaluations with multiple sites. When analyzing multiple programs simultaneously, our findings demonstrate how a Bayesian hierarchical linear model tends to shift the estimates from different programs toward each other while allowing for differences across programs or sites.

Our exploration of the Bayesian causal forest approach suggested that it can potentially be used to identify subgroups that the evaluator conducting the analysis had not specified. However, the sample sizes for the Evaluation of Employment Coaching were too small for this approach to effectively identify subgroups that had not been prespecified.

Table 1. Summary of analyses and considerations for evaluators of employment programs

Research question	Summary of results	Considerations for evaluators
How can Bayesian methods be used to reinterpret subgroup impact estimates for a single evaluation of an employment program?	• The Bayesian hierarchical linear model approach can suggest more nuanced conclusions when individual subgroup estimates are not statistically significant, especially when accounting for multiple comparisons. • Conclusions based on the Bayesian hierarchical linear model approach are less sensitive to small deviations in the data compared to conclusions based on the levels of statistical significance.	• A Bayesian hierarchical linear model can be a useful way to reinterpret subgroup impact estimates from a single evaluation. • The Bayesian hierarchical linear model approach may be especially helpful for subgroup analyses with relatively low statistical power, either because they have small sample sizes or include many different hypothesis tests.
How can Bayesian methods be used to reinterpret subgroup impact estimates for multiple evaluations of employment programs?	• The impact estimates from separate programs influenced each other, highlighting how a Bayesian hierarchical linear model draws on information across programs. • The estimated probabilities still varied across programs, showcasing how a Bayesian hierarchical linear model can allow for different conclusions for different programs.	• Using the Bayesian hierarchical linear model approach across multiple evaluations or sites can potentially increase statistical power. • Evaluators considering whether to use the Bayesian hierarchical linear model approach across multiple programs or sites may wish to carefully consider the plausibility of using different programs or sites to inform each other.
To what extent can Bayesian methods be used in an evaluation of an employment program to identify subgroups that were not previously specified?	• The Bayesian causal forest approach can potentially identify different subgroups that had not been prespecified and allow evaluators to describe results in terms of probabilities. • In many cases, the results from the Bayesian causal forest were inconsistent with the standard null hypothesis testing approach and did not suggest meaningful differences between subgroups, potentially pointing to a lack of statistical power. • Our simulation suggests that a Bayesian causal forest requires relatively large sample sizes to identify meaningful differences in impacts between subgroups, especially when exploring many subgroups.	• A Bayesian causal forest may be able to identify new subgroups, but it can also require a large sample size. • Evaluators may consider conducting a simulation to determine whether a Bayesian causal forest would be likely to detect meaningful subgroup differences. • Evaluators may benefit from selecting a limited set of candidate subgroup variables for a Bayesian causal forest to consider, especially if sample sizes are small.

I. Introduction

Many Temporary Assistance for Needy Families (TANF) recipients and other individuals with low incomes seek employment and training programs to help them find jobs or improve their earnings, which could, in turn, allow them to better support their families. However, these programs do not necessarily benefit all participants equally. In evaluations of employment programs, subgroup analysis informs whether program impacts on outcomes differ based on group characteristics (such as gender or age), suggesting whether participants with certain characteristics benefit from program participation more than others. The findings from a subgroup analysis can inform how employment programs provide their services and help them improve equity. For example, if an employment program is especially beneficial for a group that is less resourced or facing systemic barriers, then the program could focus its services more on that group. Alternatively, if a program is not working well for a group that is less resourced or facing systemic barriers, then the program could strengthen its services for that group. This report seeks to inform evaluators of employment programs about new and promising approaches to subgroup analysis.

The standard approach to subgroup analysis involves calculating separate impact estimates for groups of participants and conducting null hypothesis tests to determine whether the difference in impact estimates between the groups is statistically significant. These standard subgroup analyses are subject to three limitations. First, the results from null hypothesis tests can be hard to interpret for practitioners and policymakers (Greenland et al. 2016; Wasserstein and Lazar 2016; Gigerenzer 2018). In addition, they provide limited information on whether differences in impacts across subgroups are meaningful (Goodman 2016). Second, small changes to the analysis or data may affect whether results are statistically significant and thus change the conclusions drawn from the data (Gelman and Stern 2006; Gelman 2013). Third, impact studies are often designed to detect effects for the overall sample, but policymakers may be interested in subgroups that are only a small subset of the sample. In these cases, subgroup analyses may be underpowered. The power may be especially low when the analyses include many different subgroups, which can result in a multiple comparisons problem—the potential to find statistically significant differences by chance (Tukey 1953), which can further reduce statistical power if the problem is addressed with standard methods (Porter 2018).

To guide evaluators of employment programs, this report discusses how Bayesian methods can potentially address the limitations of standard subgroup analysis. It uses real-world data to illustrate how two Bayesian methods—a Bayesian hierarchical linear model (BHLM) and a Bayesian causal forest (BCF)—can complement the typical approach to subgroup analysis in evaluations of employment programs:

1. BHLM allows evaluators to reinterpret the original subgroup estimates in terms of probabilities rather than statistical significance, thereby providing more nuanced conclusions. For example, for an impact estimate that was not statistically significant, BHLM could still indicate whether the impact was likely to be positive. This allows evaluators to make statements such as, "There was an 80 percent chance that the impact on earnings for males exceeded $0."

2. BCF provides a flexible way to discover meaningful subgroups that had not been previously considered or specified by the evaluator conducting the analysis (Hahn et al. 2020). BCF could potentially identify subgroups for which impacts differ based on a complex combination of variables that evaluators may not consider. For example, it could reveal that an employment program conferred especially large benefits to women younger than age 27 who do not live with a child.

To explore these methods, the report uses data from four experimental evaluations of employment programs in the Evaluation of Employment Coaching for TANF and Related Populations (Evaluation of Employment Coaching) (Box I.1). This report addresses three research questions:

1. How can Bayesian methods be used to reinterpret subgroup impact estimates for a single evaluation of an employment program?
2. How can Bayesian methods be used to reinterpret subgroup impact estimates for multiple evaluations of employment programs?
3. To what extent can Bayesian methods be used in an evaluation of an employment program to identify subgroups that were not previously specified?

Our results show how the BHLM approach could complement standard null hypothesis testing for a single evaluation by accounting for multiple comparisons, providing more nuanced conclusions with small sample sizes, and guarding against small changes in the data or modeling decisions. They also show how BHLM could be used to analyze data from multiple evaluations of employment programs at once, but that this approach assumes information about subgroups in one program is relevant to other programs. Our exploration of BCF suggested that the sample sizes for the Evaluation of Employment Coaching were likely too small to identify subgroups that had not previously been specified but that BCF could be promising for studies with larger sample sizes. These findings (1) may help evaluators of employment programs better understand how Bayesian methods can be applied to subgroup analysis and consider whether these approaches fit well with their study design and (2) suggest that Bayesian methods can yield findings that better meet the needs of policymakers and practitioners in understanding which programs works best for whom.

Box I.1. Evaluation of Employment Coaching for TANF and Related Populations

To learn more about the potential of coaching to help TANF recipients and other individuals with low incomes reach economic security, the Office of Planning, Research, and Evaluation (OPRE) in the Administration for Children and Families funded an evaluation of employment coaching models. Using an experimental research design, the evaluation is examining the effectiveness and implementation of four coaching programs that aimed to help adults with low incomes succeed in the labor market. The evaluation is examining the impact of coaching on self-regulation skills and the role of self-regulation skills in generating any impacts on employment and economic outcomes.

For additional information about the evaluation and its findings, visit https://www.acf.hhs.gov/opre/research/project/evaluation-of-coaching-focused-interventions-for-hard-to-employ-tanf-clients-and-other-low-income-populations.

The rest of the report is organized as follows. In Section II, we provide an overview of the Bayesian approaches that we used to study this issue. In Section III, we describe the employment programs, data, and methods we used. In Section IV, we demonstrate how the BHLM approach can be used to reinterpret subgroup impact estimates from an evaluation of a single employment program. In Section V, we demonstrate how BHLM can be used to reinterpret subgroup estimates for multiple evaluations of employment programs. In Section VI, we explore the extent to which the BCF approach can identify new subgroups that were not previously specified by evaluators. In Section VII, we present our conclusions.

Goal4It! coach works with participant.

Photo: Rich Clement, Mathematica

II. Bayesian Approaches to Estimating Subgroup Impacts

Bayesian methods are statistical approaches that use pre-existing knowledge (called priors) about underlying relationships of interest, such as the likelihood that the impacts are of various sizes. This information can be used to improve evaluations of employment programs—that is, the impact estimates from past evaluations can inform the likely range of impacts for an evaluation of a particular program. This contrasts with standard null hypothesis testing, which does not incorporate such pre-existing knowledge about the plausible values of these relationships. Bayesian methods complement the standard reporting of hypothesis testing in three ways:

1. **Bayesian methods can generate more useful results for policymakers and practitioners.** Because Bayesian methods incorporate pre-existing knowledge, they allow researchers to draw conclusions in the form of probability statements such as, "There is an 83 percent chance that the impact exceeded $50" (Box II.1). These probabilities can complement the standard reporting of hypothesis testing by providing a more nuanced and informative conclusion than whether or not differences in impacts are statistically significant (Deke et al. 2022; Deke and Finucane 2019).

2. **Compared to standard approaches, some Bayesian methods have the potential to be relatively effective with small sample sizes.** Bayesian methods can use information from multiple subgroups or employment programs simultaneously, drawing on more information and increasing the statistical power of the analysis (Gelman et al. 2012). For example, these approaches can allow estimates from different evaluations to inform each other without assuming the impacts are identical. Additionally, because Bayesian methods provide findings on a continuum, they can be especially helpful with smaller sample sizes when subgroup analyses are at risk of being underpowered. For example, a sample size may not be large enough to detect a subgroup difference that is statistically significant at the 5 percent level using a null hypothesis test, but it may be large enough to suggest that there is an 80 percent chance that the difference is greater than zero using a Bayesian approach. The reverse situation can also arise when a small sample size yields a large and statistically significant impact estimate due to the degree of statistical noise (the component of an estimate that varies randomly). In this situation, the Bayesian approach will account for this noise and report a more moderate probability of a positive effect.

3. **Bayesian methods can address the problem of multiple comparisons in subgroup analysis.** These methods model impacts on multiple subgroups. They effectively reduce the noise in estimates, which reduces the risk that any given subgroup impact will appear remarkable just by chance and eliminates the need for post hoc multiple comparison corrections (Gelman et al. 2012; Vollmer et al. 2020).

Box II.1. Interpreting null hypothesis testing and Bayesian approaches

Null hypothesis testing and Bayesian approaches have fundamentally different interpretations, with null hypothesis testing allowing evaluators to draw conclusions about the likelihood of observing estimated impacts and Bayesian approaches allowing evaluators to draw conclusions about the likelihood of the true impacts falling into various ranges.

- **Null hypothesis testing.** The *p*-value from null hypothesis testing is the probability of finding an impact estimate at least as big as the estimated impact if the true impact were zero. For example, a *p*-value of 0.05 suggests that if the evaluation were conducted repeatedly and the true impact were zero, an evaluator would find an impact at least as large as the estimated impact only 5 percent of the time. The *p*-value is not the probability that the true impact is greater than zero. In addition, finding a statistically significant impact does not necessarily indicate that the true impact is likely near the estimated impact. The *p*-value from testing the difference in impacts between subgroups has a similar interpretation.
- **Bayesian approaches.** The probabilities from the Bayesian approaches apply to the true impact, as opposed to the estimated impact. For example, the results of a Bayesian analysis could indicate that there was a 95 percent chance that an impact was positive. Such a finding would not necessarily imply that the result was statistically significant. Conversely, a statistically significant finding may still be associated with a low probability of a true difference.

As this report illustrates, the results of a Bayesian analysis can better inform practitioners and policymakers. Probabilities from Bayesian analyses that relate to true impacts are easier to interpret and align more directly with policy questions than the probabilities from null hypothesis testing that relate to estimated impacts. For example, an evaluator might estimate a $500 difference in impacts on monthly earnings between subgroups and find that it is statistically significant at the 5 precent level. Such findings are commonly misinterpreted (Greenland et al. 2016; Wasserstein and Lazar 2016; Gigerenzer 2018). For example, one might erroneously take this finding as an indication that there was a high probability that the difference in true impacts was at least $500. Such an interpretation is natural because it is intuitive and may better align with policy decisions than the correct interpretation described above. However, because statistical significance does not inform probabilities about the true impact, the true difference could be unlikely to exceed $500. For example, a Bayesian approach using the same data could indicate that there was a 95 percent chance that the true difference in impacts was less than $100, which could have different policy implications.

This report uses two Bayesian approaches (described below) to analyze subgroups in employment program evaluations.

BAYESIAN HIERARCHICAL LINEAR MODEL

The Bayesian hierarchical linear model (BHLM) approach provides a way to reinterpret standard subgroup program impact estimates. It can be used to calculate the probability that the impact for a subgroup is positive, greater than a specified amount, or falls within certain ranges. For example, the findings from a BHLM might indicate that there was an 80 percent chance that the impact for a subgroup exceeded $50. BHLM can also provide analogous probabilities about the difference in the impacts between subgroups. For example, the analysis could indicate that there was a 70 percent chance that the impact for one subgroup exceeded that of another subgroup but a 5 percent chance that the difference exceeded $100.

Using the standard impact estimates for different subgroups, BHLM estimates how much the differences between the original subgroup estimates are due to noise versus the signal of true differences and incorporates that signal-noise breakdown to yield more precise, less noisy estimates (Lipman et al. 2022). BHLM recognizes that the impact

By drawing on multiple sources of information, a Bayesian hierarchical linear model can increase statistical power.

estimate for one subgroup provides some information about the likely impacts for other subgroups (Gelman et al. 2012). For example, the impact estimate among female participants in one employment program may inform the impact estimate for female participants in a similar employment program. It also recognizes that each impact estimate includes some noise (as measured by the standard error) and that the noisier an estimate for one program is, the less it should inform other impact estimates and, in turn, the more it should be informed by other impact estimates. This approach is sometimes called partial pooling or shrinking because it shifts the original impact estimates toward each other but still allows them to differ. As described in more detail below, compared to estimating each impact separately, partial pooling results in a set of estimates that are closer to the overall mean across the estimates. By drawing on multiple sources of information, BHLM can increase statistical power, similar to increasing the sample size or conducting a meta-analysis that combines several estimates across multiple studies.

BHLM can be used to reinterpret the impact estimates for (1) a single subgroup across multiple evaluations, (2) multiple subgroups within a single evaluation, or (3) multiple subgroups across multiple evaluations:

- **Impact estimates for a single subgroup across multiple evaluations.** BHLM can combine information across multiple evaluations to improve the impact estimates for an individual subgroup of interest (Shiferaw and Thal 2022). For example, BHLM could be used to reinterpret the impact estimates among female participants in one employment program by drawing on estimates across several evaluations of similar employment programs. In this case, BHLM would use partial pooling to shift the impact estimates for females in each program toward the average impact estimate for females across the programs. Partial pooling recognizes that, if the employment programs are comparable, several noisy estimates of a similar impact can inform each other. The amount they inform each other depends upon the noise in the original estimates, as determined by their standard errors, and estimates of the degree of signal (that is, true differences). For example, if the impact on females is estimated with more noise for one program, then that estimate will be shifted more to the estimates for the other programs. Notably, this approach still allows for each program to have a different impact on females, which contrasts with a complete pooling approach that would assume a common impact for each program.
- **Impact estimates for multiple subgroups within a single evaluation.** For a single evaluation with multiple subgroups, BHLM implements partial pooling across subgroups within the evaluation, drawing on an estimation of signal versus noise arising across the subgroups (Lipman et al. 2022). In this way, the impact estimates for various subgroups inform each other. For example, consider a scenario where BHLM is evaluating subgroups defined by gender, age, and education. If the differences in impact estimates between subgroups defined by gender and age are small and noisy, BHLM will learn from these differences that subgroups tend to have small true differences. In this case, BHLM will shift the education impact estimates more toward the overall estimate. In contrast, if the differences in impact estimates defined by gender and age are large and precisely estimated, BHLM will learn that the true subgroup differences can be large and therefore shift the education impact estimates less toward the overall impact.

- **Impact estimates for multiple subgroups across multiple evaluations.** For multiple evaluations and subgroups, BHLM implements partial pooling on two dimensions: (1) across evaluations for each individual subgroup and (2) across subgroups within an evaluation (Shiferaw and Thal 2022).

Because we expect that most evaluations will have multiple subgroups, this report explores the latter two applications: (1) partial pooling of multiple subgroups within an evaluation and (2) partial pooling of multiple subgroups across multiple evaluations.

In each of these applications, BHLM uses the original impact estimates to disentangle the signal from the noise across subgroups and/or across evaluations. The standard errors of the impact estimates provide a measure of the noise for individual impact estimates. BHLM also requires information on the correlation of the subgroup estimates due to the overlapping samples for subgroups. For example, to the extent that younger participants tend to have less education, impact estimates for younger and less educated subgroups will be correlated because the overlapping sample of young, less educated participants will contribute to both estimates. These correlations can be calculated with seemingly unrelated estimation (Appendix A). BHLM calculates the degree of signal from the differences between the impact estimates in the evaluation and prior evidence on the degree of true differences between subgroups or evaluations from a meta-analysis of evaluations in similar settings.

As a Bayesian model, BHLM relies on prior evidence about the magnitude of various relationships. Specifically, BHLM requires prior evidence about the likelihood of the magnitude of (1) impacts, (2) differences in impacts between subgroups, (3) differences in impacts between evaluations, and (4) differences in impacts due to unobserved sources that are not explained by the subgroups and evaluation. This prior evidence can be estimated from a meta-analysis of similar evaluations.

The benefits of partial pooling can be substantial when dealing with small subgroups, which makes BHLM especially useful when sample sizes for subgroups of interest are limited. BHLM also naturally corrects for issues of multiple comparisons by moving noisier estimates further toward the overall mean through partial pooling. In addition, because BHLM builds on standard impact estimates, it allows for direct comparison to the standard estimates and requires limited additional computational time (see Appendix A for more details on the procedure).

As with any modeling approach, these benefits rely on the underlying assumptions of the model being met. For BHLM, these key assumptions are (1) that the prior evidence pertains to the evaluation of interest; (2) that the model captures the process that determines the effectiveness of the interventions (for example, there are no unmeasured subgroups that relate to effectiveness); and (3) that pooling across evaluations is sensible (for example, there are underlying shared effects between the different evaluations, so they ought to inform one another). Because these assumptions are not testable, we recommend that evaluators consider other sources of evidence and also conduct sensitivity analyses. For example, if an evaluator is deciding whether or not to use partial pooling across multiple evaluations, we suggest considering the programs' logic models and conducting sensitivity analyses without pooling across evaluations.

We selected BHLM for subgroup analysis of employment programs over alternative approaches because it (1) draws on the results from standard impact analyses, ensuring that the BHLM results are consistent with them; (2) accounts for multiple comparisons (Gelman et al. 2012); and (3) provides a framework for combining estimates across multiple programs or multiple program sites.

BAYESIAN CAUSAL FOREST

The Bayesian causal forest approach can potentially identify subgroups that had not been previously specified by the evaluator conducting the analysis.

The Bayesian causal forest (BCF) approach can potentially identify subgroups that had not been previously specified by the evaluator conducting the analysis (Hahn et al. 2020). Like the BHLM approach, BCF also addresses the multiple comparisons problem by shifting estimated impacts for different subgroups toward each other (Hahn et al. 2020). Unlike the BHLM approach, BCF does not draw on standard impact estimates, so it does not provide a way to reinterpret them.

BCF involves re-estimating impacts as flexible functions of variables that define potential subgroups. This flexibility allows BCF to search for subgroups with large differences in impacts that evaluators did not prespecify. For example, BCF could find that impacts not only vary by gender or race, but also by combinations of gender and race. BCF can also identify subgroups based on cutoffs in continuous measures that were not prespecified. For example, an evaluator could allow BCF to explore whether impacts differ by monthly earnings but not prespecify specific earnings categories. BCF might identify substantial impacts for those earning less than $400 per month or more than $1,500 per month. These features make BCF appealing when an evaluator lacks a theoretical basis for which subgroups may matter and seeks to explore the possibilities.

Rather than starting from standard estimates, BCF fits an entirely new model from the ground up (Hahn et al. 2020). The BCF approach—sometimes referred to as a "black box" approach—is flexible and nonparametric, which means that it relies on few assumptions about the relationship between variables. BCF works by repeatedly forming subgroups in a variety of flexible ways and examining whether impacts differ across those subgroups. If it found substantial evidence that impacts differed for a potential subgroup, then its final estimates would more likely account for differences by that subgroup.

BCF also uses Bayesian techniques to avoid overfitting—that is, to keep from finding patterns in the data that do not generalize, which is an issue with other similar models (for example, the random forest approach).[1] To achieve this, BCF builds its model iteratively by exploring new potential subgroups one at a time (for example, allowing for differences by gender, then allowing gender differences to vary by race; or allowing for an increasing number of breakpoints in the relationship between prior earnings and outcomes). BCF requires increasingly more evidence—measured as an improvement in how well the model fits the data—to identify more complicated subgroups (for example, those that depend upon many different variables).

[1] In the random forest approach, a host of individual decision trees are each fit to a random subsample of data using a random subset of covariates. In each decision tree, the sample is iteratively split by covariate values according to which splits best fit the outcome being modeled. For example, a tree might first separate the sample by gender, and then may further split males by age. In contrast to BCF, the random forest approach does not prioritize simpler models over more complex ones, nor does it shift estimates for smaller groups closer to the overall mean.

BCF also employs a form of partial pooling, like BHLM, by shifting the estimated impacts for individual groups back toward the overall estimate in proportion to the precision of the subgroup differences, largely driven by sample size. For example, if there were very few female participants, then the estimate for those participants would be shifted closer to the overall estimate. Importantly, BCF's Bayesian priors and avoidance of overfitting address the multiple comparisons problem by allowing simultaneous testing for many newly identified subgroups.

One drawback to BCF is that it repeats the process of fitting a model to the data thousands of times, meaning that it does not directly identify a single set of subgroups or control covariates. Instead, the BCF approach yields estimated impacts for each individual in the data. To identify particular subgroups for this report, we used a second model, classification and regression trees (CART) (Breiman et al. 1984), which examines the estimated impacts across individuals in the data to determine which predictors form the most important subgroups (see Appendix A for additional details).

Another related drawback is that BCF cannot reinterpret standard impact estimates, unlike BHLM. Like BHLM, BCF could be used to estimate the probability that impacts for prespecified subgroups were below or above certain thresholds. However, BCF estimates these probabilities based on a completely separate impact model that may not align with the standard estimates, which makes apples-to-apples comparisons more difficult. Comparing the results between BCF and the original impact analysis is like comparing two different models that include different covariates. In an uncompromised experimental evaluation, the two models should yield similar results on average; in practice, they will differ from each other. In the case of BCF, this comparison is even more challenging because BCF fits the model repeatedly and does not produce a definitive set of subgroups or covariate relationships. In contrast, BHLM's estimates are based on the standard impact estimates, which ensures results that are aligned with those estimates.

To ensure that the subgroups that BCF identified as meaningful did not arise from chance differences in the modeling approaches, we also tested these subgroups by using the original impact estimation model. We placed greater weight on findings that aligned between the two approaches. More generally, in cases where the BCF estimates and the original impact estimates differ substantially, evaluators may examine the BCF results to see how the BCF model differs from the original impact model. For example, the BCF approach allows for nonlinear relationships between covariates and outcomes that are typically not included in standard impact estimates. Comparing the two models could help evaluators locate the source of the differences.

Despite its drawbacks, we selected the BCF approach because, compared to other similar approaches that provide flexible ways to estimate causal impacts with few parametric assumptions, BCF has been effective at estimating the true impacts while limiting the false discovery of subgroup impacts (Dori et al. 2019; Hahn et al. 2020; Thal and Finucane 2023).

III. Overview of Programs, Data, and Methods

To illustrate how Bayesian methods can apply to real-world evaluations of employment programs, we reanalyzed data from the Evaluation of Employment Coaching, which includes evaluations of four separate employment coaching programs. This section describes the programs, the design of the original impact evaluations and data collected, and the subgroup methods used for this report.

EMPLOYMENT COACHING PROGRAMS

In employment coaching trained staff work collaboratively with participants to help them set individualized goals that are directly or indirectly related to employment, and then they provide motivation, support, and feedback as participants work toward those goals (Joyce and McConnell 2019). Unlike most traditional case managers, coaches work as partners to participants and thus are not directive—they do not tell the participants what goals they should pursue or what action steps to take in pursuing them.

The Evaluation of Employment Coaching tested the effectiveness of four employment coaching programs (described here as implemented during the study):

1. **Family Development and Self-Sufficiency program (FaDSS) in Iowa.** Under contract to the state, 17 local human services agencies used grants from the Iowa Department of Human Rights to provide Temporary Assistance for Needy Families (TANF) recipients with coaching during home visits. Seven of those 17 agencies participated in the evaluation.

2. **Goal4 It!™ in Jefferson County, Colorado.** Goal4 It! was an employment coaching program designed by Mathematica and its partners that was piloted in a TANF program as an alternative to more traditional case management.

3. **LIFT in Chicago, Los Angeles, New York City, and Washington, DC.** LIFT is a nonprofit organization that provides career and financial coaching to parents and caregivers of young children. LIFT sites in Chicago, Los Angeles, and New York City participated in the evaluation.

4. **MyGoals for Employment Success in Baltimore and Houston.** MyGoals was a coaching demonstration project designed by MDRC and its partners that provided employment coaching and incentives to unemployed adults receiving housing assistance. It was operated within the Housing Authority of Baltimore City and the Houston Housing Authority.

Even though all the programs served adults with low incomes, programmatic variation in eligibility criteria, settings, and available services may have contributed to differences in characteristics of the study participants across the programs (Table 2). For example, compared with the other three programs, LIFT served a much higher proportion of Hispanic participants. In addition, 38 percent of participants in the LIFT program

did not have a high school diploma or GED certificate, compared to 22 percent to 25 percent of participants in the other programs. Among the three programs with monthly self-reported earnings for the 30 days prior to study enrollment, participants' earnings ranged from $160 to $624 when including participants who did and did not work. Participants in MyGoals were the oldest, with an average age of 38, while those in FaDSS were the youngest, with an average age of 29.

Table 2. Baseline characteristics of study participants

Baseline characteristic	FaDSS	Goal4 It!	LIFT	MyGoals
Demographics				
Average age (in years)	29	32	33	38
Female (percentage)	94	90	95	88
Race and ethnicity (percentage)				
Hispanic	12	42	71	3
Black, non-Hispanic	36	9	28	95
White, non-Hispanic	48	47	1	2
Other	3	3	1	1
Currently married (percentage)	7	12	35	NA
Number of children respondent lives with	2.1	1.9	2.3	1.6
Socioeconomic status				
Does not have high school diploma or GED (percentage)	24	22	38	25
Receiving public assistance (percentage)	99	93	84	100
Worked for pay in past 30 days (percentage)	34	27	52	NA
Self-reported earnings in past 30 days ($)				
All study participants	161	160	624	NA
Among those who worked for pay in past 30 days	481	601	1,195	NA
Worked for pay in past quarter (NDNH; percentage)	58	49	NA	35
Average monthly earnings reported to a UI agency in the past quarter (NDNH; $)				
All study participants	498	733	NA	340
Among those with positive earnings reported to a UI agency	864	1,491	NA	980
Sample size	863	802	807	1,799

Source: Evaluation of Employment Coaching baseline survey, MyGoals Baseline Questionnaire data, public housing agency administrative data, and the National Directory of New Hires.

Note: Baseline characteristics were drawn from the baseline survey unless otherwise noted.

NA = not available; NDNH = National Directory of New Hires; UI =Unemployment Insurance.

IMPACT EVALUATION DESIGN AND DATA SOURCES

The original evaluation used a random assignment design. Adults who were eligible for coaching and who consented to participate in the evaluation were randomly assigned to either a program group that was eligible to receive the program's coaching services or a control group that was not eligible for such services. In total, 4,276 adults who were eligible for one of the four employment coaching programs and who consented to participate in the study were randomly assigned in this way. The number of participants ranged from 802 for Goal4 It! to 1,803 for MyGoals.[2]

Baseline data on participant characteristics were collected during a baseline survey or, for MyGoals, a baseline questionnaire administered to study participants at the time of study enrollment.[3] In this report, we primarily used these baseline data to form subgroups.

The original evaluation assessed each program's impacts on participants' self-regulation, employment, earnings, self-sufficiency, and other measures of well-being. The evaluation collected data on these outcomes through several surveys and administrative records. The first follow-up survey occurred at 9 months (for FaDSS, Goal4 It!, and LIFT) or 12 months (for MyGoals) following random assignment. Second and third follow-up surveys occurred at 21 months and 48 months, respectively. This report focuses on impacts on monthly earnings as self-reported on the first follow-up survey.

The original evaluation presented standard impact estimates based on null hypothesis testing, as well as a complementary Bayesian analysis for selected outcomes on the full sample (as opposed to subgroups). The original Bayesian analysis used prior evidence from a meta-analysis of similar evaluations from the Pathways to Work Evidence Clearinghouse,[4] as described in Shiferaw and Thal (2022) and Moore et al. (2023). For the BHLM analysis presented in this study, we used the same prior evidence. See Appendix A for more details.

METHODS FOR SUBGROUP ANALYSIS

Defining subgroups

For this study, we selected a set of subgroups with the goal of illustrating how the BHLM and BCF approaches work in practice for evaluations of employment programs. For the BHLM approach, we selected subgroups based on those analyzed in the Evaluation of Employment Coaching (Moore et al. 2023) to ensure that they would be relevant to other evaluations of employment programs. To limit the multiple comparisons problem, the original evaluation conducted subgroup analyses with a selected set of subgroups. We examine those same subgroups, which were defined by: (1) age, (2) number of children, (3) education level, (4) race and ethnicity, (5) goal-setting skills, (6) recent

[2] The number of participants randomly assigned for MyGoals does not match the sample size in Table 2 because four participants withdrew from the study.

[3] The available baseline data differed by study program because of differences in the study intake process. Study intake for the MyGoals program began before the baseline data collection instruments for the other programs were developed and approved.

[4] The Pathways to Work Evidence Clearinghouse provides evidence on interventions aimed at improving employment outcomes for individuals with low incomes. More information can be found at: https://pathwaystowork.acf.hhs.gov/.

employment status, (7) barriers to employment, (8) having a valid driver's license, (9) preferred language, (10) program location, and (11) urbanicity. To better demonstrate how BHLM can address the multiple comparisons problem, we also included some additional subgroups beyond those in the original evaluation. In particular, when available, we used the following variables to define subgroups for each program: (1) the extent to which transportation made it hard to find employment, (2) the extent to which lack of child care made it hard to find employment, and (3) disability status. In the body of this report, we present selected estimates that illustrate the performance of the methods. Appendix B includes the full set of estimates. For the BCF approach, we sought to explore how it could potentially identify new subgroups based on a broader set of variables, so we examined all available baseline variables for each program.

Standard subgroup impact estimation methods based on null hypothesis testing

To provide a benchmark for interpreting the Bayesian subgroup analysis, we also conducted a standard subgroup impact analysis based on the null hypothesis testing framework, following the approach used in the Evaluation of Employment Coaching (Moore et al. 2023). We first defined two subgroups using the variables described above. For example, for participant age, we defined subgroups for whether someone was (1) older than 30 years old or (2) 30 years old or younger. Then, for each program, we estimated subgroup impacts from a model that pooled data for the two subgroups and included indicators for the research group (program or control) and the subgroups as well as an interaction between the research group indicator and the subgroup indicator. To improve precision and account for baseline differences between members of the program and control groups, we included the covariates used in the original evaluation (Moore et al. 2023) (see Appendix A for a list of covariates). We conducted statistical tests for whether individual subgroup impacts were greater than zero as well as tests for the differences between the two subgroup estimates. See Appendix A for more details on this approach.

IV. How Can Bayesian Methods Be Used to Reinterpret Subgroup Impact Estimates for a Single Evaluation of an Employment Program?

In this section, we use the BHLM approach to reinterpret subgroup impact estimates from a single evaluation of an employment coaching program. We present probabilities that the impacts are above or below particular thresholds, as opposed to statistical significance. BHLM can account for the multiple comparisons problem and provide more nuanced information on the probability that subgroups differ compared to standard null hypothesis testing, which can help policymakers and practitioners better serve and design programs for specific groups.

APPROACH

Using the BHLM approach described in Section II, we reinterpreted standard subgroup impact estimates for Goal4 It!, one of the four programs in the Evaluation of Employment Coaching. We present the Goal4 It! analyses in this report because the findings illustrate some of the key features of the BHLM approach. We estimated impacts on self-reported monthly earnings and formed subgroups based on 10 binary variables that resulted in a total of 20 different subgroups. Our discussion focuses on subgroups based on four variables that best show how BHLM performs using our data: (1) participant age, (2) number of children, (3) education level, and (4) race and ethnicity. See Appendix B for findings based on all 10 variables as well as comparable findings for the other three programs.

We present findings from the following two analyses:

1. **Null hypothesis testing.** We provided impact estimates and levels of statistical significance using a standard null hypothesis testing framework. To address the multiple comparisons problem, we used the Benjamini-Hochberg procedure to adjust the levels of significance (Benjamini and Hochberg 1995). This procedure determines whether the impact difference between two subgroups is significant after accounting for the total number of hypothesis tests.

2. **BHLM approach.** Using BHLM, we estimated the probability that the true impacts were above and below various thresholds and the comparable probabilities for the difference in impacts between subgroups.

RESULTS

The results from the null hypothesis testing and the BHLM approaches were consistent.

In cases where individual subgroup impact estimates were higher, the BHLM approach indicated that the impact estimates were more likely to be positive (Table 3). For

example, of the 20 subgroups, the impact estimate on self-reported monthly earnings was greatest for participants who were age 30 or younger ($351). Similarly, the BHLM approach revealed that there was an 87 percent chance that the impact for that subgroup exceeded zero, the highest probability across all subgroups. In addition, when the estimated difference in impacts between subgroups was greater, the BHLM approach indicated that differences were more likely to be higher. For example, the largest difference in impacts between the subgroups was for participants who were age 30 or younger versus those who were older than age 30. The probability that this difference exceeded zero was also the highest at 74 percent.

The BHLM approach can suggest more nuanced conclusions when individual subgroup estimates are not statistically significant.

Only two of the impact estimates for individual subgroups were statistically significant at the 5 percent level: (1) those age 30 or younger and (2) those with fewer than two kids. Using the null hypothesis testing framework, an evaluator might report that the program did not have statistically significant impacts for any other subgroups and focus the conclusions on the two statistically significant cases. However, the BHLM approach paints a more nuanced picture. For example, the impact estimate for participants with no college was positive but only statistically significant at the 10 percent level. In contrast, the BHLM approach indicated that there was an 83 percent chance that the impact was positive for that subgroup, suggesting that the program likely improved self-reported monthly earnings for that subgroup. However, the BHLM approach also indicated that there was only a 28 percent chance that this impact exceeded $50, suggesting that it was likely small. Even though the BHLM approach suggests that the impact was relatively small, this information could still be useful to practitioners. For example, if implementing the employment program were relatively low cost, then the relatively small benefits could outweigh the costs for that subgroup.

The results highlight the importance of accounting for the multiple comparisons problem and how the BHLM approach can help draw more nuanced conclusions about the differences between subgroups.

Before accounting for multiple comparisons, two differences in impacts between subgroups were statistically significant at the 5 percent level: those defined by age and number of children. However, after accounting for multiple comparisons, these two differences were no longer statistically significant. Therefore, based on null hypothesis testing, an evaluator might suggest that the impacts on self-reported earnings did not differ across subgroups. However, the results from the BHLM approach pointed to more nuanced conclusions. For example, the results indicated that there was a 74 percent chance that the impact for participants who were age 30 or younger exceeded the impact for those who were older than age 30, suggesting that it was likely that younger participants benefited more. At the same time, there was only a 23 percent chance that the difference exceeded $50, suggesting that the difference was likely small.

Conclusions based on the BHLM approach are less sensitive to small deviations in the data compared to conclusions based on the levels of statistical significance.

Although none of the differences in impacts on self-reported monthly earnings across subgroups were statistically significant after adjusting for multiple comparisons, a very small deviation in the data could have radically changed conclusions based on the null hypothesis testing approach. If the unadjusted *p*-value for the estimated difference in impacts between subgroups defined by the number of children were 0.00009 lower, then after correcting for multiple comparisons the differences in impact estimates would have been statistically significant for subgroups defined by age and number of children. This lower *p*-value could be achieved if a single participant in the program group reported an additional $50 of monthly earnings. With this small difference, an evaluator using the null hypothesis framework might have concluded that the program had large, statistically significant impacts on self-reported monthly earnings based on age and number of children and focused the conclusions on those subgroups. In contrast, for these subgroups, BHLM's estimate of the probability that the difference between impacts was greater than zero did not change by more than a percentage point. An evaluator using the BHLM approach would have drawn a similar conclusion before and after the change.

The results demonstrate how the BHLM approach can suggest more interpretable conclusions about the differences between subgroup impacts, even in cases where the differences are statistically significant.

The Bayesian hierarchical linear model results provide a fundamentally different way of interpreting the difference in impact estimates compared to the null hypothesis testing framework.

The BHLM results provide a fundamentally different way of interpreting the difference in impact estimates compared to the null hypothesis testing framework. For example, with the small deviation in the data described above, the estimated difference in impacts on monthly earnings for participants who were age 30 or younger versus older than age 30 was statistically significant at the 5 percent level and large in magnitude ($497). It is common to misinterpret such findings as suggesting that there was a high probability that the true difference was near the estimated difference. From this result, some might conclude that the true difference in impacts on monthly earnings was likely large (around $500), suggesting that the employment program might wish to strengthen its services and supports for older participants. However, the correct interpretation is, if the true difference in impacts were $0, the estimated impact would reach $497 by chance less than 5 percent of the time. The BHLM results highlight how this statement does not imply that the true difference was likely close to $500. The results indicate that it was more likely that younger participants benefited more than older participants. However, they also suggest that (1) there was only a 4 percent chance that this difference exceeded $100 and (2) that it was most likely less than $25. With this interpretation, the implications for an employment program could be very different.

Table 3. Impacts of a single coaching program (Goal4 It!) on average monthly self-reported earnings by subgroup based on the BHLM approach

				Probability that the true impact or difference in true impacts was...							
Subgroup	Sample size	Control group mean	Estimated impact or difference and level of significance	Less than -$100	Less than -$50	Less than -$25	Less than $0	Greater than $0	Greater than $25	Greater than $50	Greater than $100
Participant age											
Age 30 or younger	188	452	351***	0.00	0.00	0.02	0.13	0.87	0.65	0.38	0.07
Older than age 30	251	983	-146	0.00	0.01	0.07	0.30	0.70	0.36	0.13	0.01
Difference between subgroups			497***	0.00	0.01	0.07	0.26	0.74	0.47	0.23	0.04
Number of children											
Fewer than two	194	504	347**	0.00	0.00	0.04	0.20	0.80	0.52	0.24	0.03
Two or more	242	941	-131	0.00	0.01	0.07	0.30	0.70	0.38	0.14	0.01
Difference between subgroups			478**	0.00	0.02	0.11	0.35	0.65	0.31	0.10	0.01
Education level											
No college	216	478	193*	0.00	0.00	0.03	0.17	0.83	0.55	0.28	0.03
Some college or higher	223	964	61	0.00	0.00	0.04	0.21	0.79	0.50	0.24	0.03
Difference between subgroups			132	0.01	0.07	0.20	0.45	0.55	0.27	0.10	0.01
Race and ethnicity											
Not Hispanic/ Latino/a	208	714	146	0.00	0.00	0.04	0.19	0.81	0.55	0.29	0.05
Hispanic/ Latino/a	145	737	109	0.00	0.00	0.04	0.18	0.82	0.56	0.31	0.08
Difference between subgroups			37	0.04	0.13	0.27	0.52	0.48	0.24	0.10	0.02

Sources: Evaluation of Employment Coaching baseline and first follow-up surveys

Note: Outcomes are measured over the first 9 months after study enrollment. Estimated impacts and level of statistical significance come from the standard approach to estimating subgroup impacts based on the null hypothesis testing framework.

***/**/* Impact estimates are statistically significant at the .01/.05/.10 levels, respectively, using a two-tailed t-test.

DISCUSSION

These analyses showed how the BHLM approach can complement the null hypothesis testing approach in subgroup analyses of a single employment program. The findings highlighted how the results from the BHLM approach were generally consistent with the null hypothesis testing approach but that the BHLM approach had the potential to provide more nuanced conclusions that were less sensitive to small variations in the data. For these reasons, the BHLM approach may be especially helpful for subgroup analyses with relatively low statistical power, either because they have small sample sizes or include many different hypothesis tests.

My Goals coach works with participant

Photo: Rich Clement, Mathematica

V. How Can Bayesian Methods Be Used to Reinterpret Subgroup Impact Estimates for Multiple Evaluations of Employment Programs?

The BHLM approach also provides a way to combine information across multiple evaluations of similar programs or a single evaluation with multiple sites. For example, it could be used to estimate the impact of similar programs among participants in a particular subgroup. BHLM uses the impact estimate for the subgroup in each program to inform the likely estimate for the same subgroup in the other programs (Gelman et al. 2012). It also recognizes that the impact estimate for the subgroup in each program includes some uncertainty and that the more uncertainty there is in an estimate for one program, the less it should inform those of other programs.

APPROACH

We explored how BHLM could be used to reinterpret subgroup impact estimates across the four employment coaching programs in the Evaluation of Employment Coaching. We compared the results from two models: (1) one that used BHLM separately with data from each of the four programs and (2) one that used BHLM to combine information across all four programs. As with the analysis in Section IV, we estimated impacts on self-reported monthly earnings and formed subgroups based on binary variables. In this section, we focus our discussion on one illustrative subgroup defined by whether participants were age 30 or younger versus older than age 30. Appendix B presents comparable findings for the other subgroups that were common across all programs.

RESULTS

When analyzing each program separately, the estimated impacts varied substantially across programs, along with the probability that the impacts exceeded various thresholds.

The estimated impacts on self-reported monthly earnings for those who were age 30 or younger were especially variable, with positive and statistically significant impacts for Goal4 It!, a positive but insignificant impact for FaDSS and LIFT, and a negative but insignificant impact for MyGoals (Panel [a] of Table 4). However, the findings from the two approaches were consistent. The corresponding probabilities that the impacts were positive mirrored the impact estimates, ranging from 66 percent (MyGoals) to 87 percent (Goal4 It!).

When using BHLM with all four programs simultaneously, the impact estimates from separate programs influenced each other—highlighting how BHLM draws on information across programs.

On average, across all programs, the estimated impact on self-reported monthly earnings was positive for those who were age 30 or younger. For this reason, when using BHLM with all programs simultaneously, the probabilities that the corresponding impacts were positive increased compared to the results when estimating the probabilities separately by program. The impacts all shifted toward this more positive value. The difference was especially striking for MyGoals. When using BHLM on MyGoals alone, there was a 66 percent chance that MyGoals had a positive impact on self-reported monthly earnings for this subgroup (Panel [b] of Table 4).[5] When estimating BHLM for all programs simultaneously, the chance that MyGoals had a positive impact increased to 82 percent because the analysis accounted for the information from the other programs, all of which had positive impacts.

When using BHLM with all programs simultaneously, the estimated probabilities still varied across programs, which showcased how BHLM allows for different conclusions for different programs.

After using BHLM for all programs simultaneously, the probability that the programs had a positive impact on self-reported monthly earnings for those age 30 or younger ranged from 82 percent to 92 percent, which was much narrower than the range of 66 percent to 87 percent when analyzing each program separately (Table 4). This difference in range reflects how BHLM shifts estimates closer together when using partial pooling across programs. However, when considering the probabilities at other thresholds, the conclusions still differed across programs when using all programs simultaneously. For example, there was a 44 percent chance that the impact of Goal4 It! exceeded $50, while there was only a 20 percent chance that the impact of MyGoals exceeded $50. Therefore, an evaluator could still draw different conclusions about the impacts across the programs.

[5] Even though the estimated impact for MyGoals was negative, the probability exceeded 50 percent because (1) the prior was slightly positive and (2) MyGoals overall had a positive (but insignificant) impact on survey-reported earnings in the full sample.

Table 4. Impacts of all four coaching programs on average monthly self-reported earnings by participant age at baseline based on the BHLM approach

					Probability that the true impact or difference in true impacts was...							
Program	Subgroup	Sample size	Control group mean	Estimated impact or difference and level of significance	Less than -$100	Less than -$50	Less than -$25	Less than $0	Greater than $0	Greater than $25	Greater than $50	Greater than $100
(a) Estimating separately for each program												
FaDSS	Age 30 or younger	289	559	199**	0.00	0.00	0.02	0.16	0.84	0.53	0.22	0.01
	Older than age 30	208	737	-66	0.00	0.01	0.08	0.32	0.68	0.36	0.13	0.01
	Difference between subgroups			265*	0.00	0.02	0.11	0.32	0.68	0.35	0.11	0.00
Goal4 It!	Age 30 or younger	188	452	351***	0.00	0.00	0.02	0.13	0.87	0.65	0.38	0.07
	Older than age 30	251	983	-146	0.00	0.01	0.07	0.30	0.70	0.36	0.13	0.01
	Difference between subgroups			497***	0.00	0.01	0.07	0.26	0.74	0.47	0.23	0.04
LIFT	Age 30 or younger	198	790	36	0.00	0.01	0.09	0.33	0.67	0.33	0.11	0.00
	Older than age 30	376	901	13	0.00	0.00	0.07	0.34	0.66	0.30	0.08	0.00
	Difference between subgroups			23	0.00	0.04	0.18	0.49	0.51	0.20	0.05	0.00
MyGoals	Age 30 or younger	289	357	-34	0.00	0.01	0.09	0.34	0.66	0.32	0.10	0.00
	Older than age 30	962	386	56	0.00	0.00	0.02	0.22	0.78	0.34	0.07	0.00
	Difference between subgroups			-90	0.00	0.04	0.22	0.58	0.42	0.14	0.03	0.00
(b) Combining information across programs												
FaDSS	Age 30 or younger	289	559	199**	0.00	0.00	0.01	0.10	0.90	0.63	0.28	0.02
	Older than age 30	208	737	-66	0.00	0.00	0.05	0.23	0.77	0.44	0.17	0.01
	Difference between subgroups			265*	0.00	0.02	0.10	0.32	0.68	0.34	0.10	0.00
Goal4 It!	Age 30 or younger	188	452	351***	0.00	0.00	0.01	0.08	0.92	0.72	0.44	0.09
	Older than age 30	251	983	-146	0.00	0.00	0.04	0.23	0.77	0.45	0.17	0.01
	Difference between subgroups			497***	0.00	0.01	0.07	0.25	0.75	0.47	0.22	0.04
LIFT	Age 30 or younger	198	790	36	0.00	0.00	0.03	0.17	0.83	0.53	0.22	0.01
	Older than age 30	376	901	13	0.00	0.00	0.04	0.24	0.76	0.38	0.11	0.00
	Difference between subgroups			23	0.00	0.02	0.11	0.37	0.63	0.29	0.08	0.00
MyGoals	Age 30 or younger	289	357	-34	0.00	0.00	0.03	0.18	0.82	0.51	0.20	0.01
	Older than age 30	962	386	56	0.00	0.00	0.01	0.16	0.84	0.43	0.11	0.00
	Difference between subgroups			-90	0.00	0.02	0.12	0.44	0.56	0.21	0.05	0.00

Sources: Evaluation of Employment Coaching baseline and first follow-up surveys and MyGoals Baseline Questionnaire data.

Note: Outcomes are measured over the first 9 months after study enrollment for FaDSS, Goal4 It!, and LIFT. Outcomes are measured over the first 12 months after study enrollment for MyGoals. Estimated impacts and level of statistical significance come from the standard approach to estimating subgroup impacts based on the null hypothesis testing framework.

***/**/* Impact estimates are statistically significant at the .01/.05/.10 levels, respectively, using a two-tailed t-test.

DISCUSSION

A Bayesian hierarchical linear model may present an attractive approach in cases where evaluators expect some variation in impacts across programs or sites.

We explored using BHLM to reinterpret impact estimates from multiple programs simultaneously. The results highlight how this approach represents a middle ground between analyzing programs or sites separately from each other and pooling the data across programs or sites to estimate a common impact. In contrast to analyzing each program separately, BHLM combined information across programs by shifting the impact estimates toward each other. In contrast to estimating a common impact, BHLM allowed for the impacts to differ across programs. Therefore, BHLM may present an attractive approach in cases where evaluators expect some variation in impacts across programs or sites, but the programs or sites are similar enough that they can plausibly inform each other.

However, BHLM may not work well for all applications with multiple evaluations or sites. BHLM assumes that the information about subgroup impacts from one program can meaningfully inform those of other programs, but that assumption may not hold true in practice. In the Evaluation of Employment Coaching, the programs were substantively different from each other (Moore et al. 2023), so combining information across programs could yield misleading results. For example, our analyses showed that, compared to using BHLM separately by program, using data from all programs simultaneously led to a different conclusion about the impacts by age for one of the programs (MyGoals). Without additional information, it is difficult to tell whether this result was justified. On the one hand, the estimate for MyGoals may have been a noisy outlier, so adding information from the other programs could have reduced the noise. On the other hand, the employment coaching provided through MyGoals differed from the other programs. For example, unlike in other programs, MyGoals coaches used a questionnaire that assessed participants' strengths and weaknesses in self-regulation skills and discussed these skills explicitly with participants. For this reason, combining information in this way may not have been warranted.

Evaluators considering this partial pooling approach may wish to carefully consider the plausibility of using different programs or sites to inform each other. For example, evaluators might consider other information, such as the similarity between the logic models or services across the programs or information about program staff's experiences with serving different subgroups. As with other subjective modeling decisions, evaluators may benefit from conducting sensitivity tests that examine how partial pooling compares to estimating impacts separately by program or site.

VI. To What Extent Can Bayesian Methods Be Used in an Evaluation of an Employment Program to Identify Subgroups That Were Not Previously Specified?

The BCF approach provides a way to identify previously unspecified subgroups for which impacts vary. In standard subgroup analyses, evaluators must prespecify subgroups of interest, often drawing on theories about which participants may benefit more than others. For this reason, evaluators often examine a limited number of subgroups, rather than considering complex combinations of multiple subgroups. Additionally, for continuous variables such as age, evaluators must choose a cutoff to divide the subgroup, a choice that typically does not depend upon the estimated impacts above or below that cutoff. For this reason, the standard approach may overlook subgroups that experience a significantly different impact based on other cutoffs.

BCF allows researchers to include complex combinations of subgroups while also correcting for multiple comparisons. It also uses the data to discover new cutoffs when defining subgroups based on continuous variables. In addition, BCF also provides a way to present the differences between subgroups in terms of probabilities rather than statistical significance. However, using BCF to identify new subgroups may require larger sample sizes than those available in many employment evaluations.

APPROACH

Following the method outlined in Section II, we used the BCF approach to estimate an impact on self-reported monthly earnings for each individual participant, allowing the model to consider all baseline data in the Evaluation of Employment Coaching. We estimated these impacts separately by employment program. The model allowed the impacts to differ based on complex combinations of subgroups. For each of the four employment programs, we then applied CART to identify the top subgroup variable—the one with the largest difference in estimated impacts between the two subgroups. For comparison, we also estimated impacts and levels of statistical significance for these subgroups using a standard null hypothesis testing approach.

To explore the statistical power of the BCF approach, we conducted simulations using different sample sizes, outcome distributions, and numbers of candidate subgroup variables (the variables that BCF considers when defining subgroups). We simulated data that assumed a true difference in impacts of 0.20 standard deviations for one subgroup and no difference in impacts between all other subgroups. The 0.20 standard deviation difference translated to a difference in earnings between groups of $200, in line with what we estimated for some subgroups using the standard impact estimation approach. We then estimated the BCF on the simulated data and determined whether there was a likely subgroup difference, which we defined as whether there was a 95 percent chance that the impact estimated for one subgroup was larger than that of the other.

For each sample size and number of candidate subgroup variables, we simulated 100 different data sets and repeated this calculation on each. We then calculated the percentage of times that the model estimated that there was a likely subgroup difference. This percentage is an estimate of statistical power. For example, if the estimate were 80 percent, the finding would have the following interpretation: in the case of a true difference in subgroup impacts of 0.20 standard deviations, BCF would estimate that there was a high likelihood that the subgroups differed 80 percent of the time.

RESULTS

The results highlighted that the BCF approach can potentially identify different subgroups that had not been prespecified and allow evaluators to describe results in terms of probabilities.

The top subgroup variables identified through the BCF were based on (1) any work in the last 30 days, (2) whether the baseline survey was conducted in Spanish, and (3) being a single parent (Table 5). Like the BHLM approach, the BCF approach suggested that many of the impacts between subgroups were likely positive, even when the null hypothesis testing approach did not suggest that they were statistically significant. For example, the impact estimate for FaDSS was positive but not statistically significant among those without recent employment. However, BCF indicated that there was an 81 percent chance that the impacts were positive for that subgroup.

The BCF estimates and the standard impact estimates did not always align, highlighting why BCF cannot be used to reinterpret standard impact estimates.

We recommend putting more weight on subgroups that both the Bayesian causal forest approach and the original impact estimation reveal as meaningful.

For example, across the four programs, the standard subgroup impacts on self-reported monthly earnings differed the most by employment status for Goal4 It!, with greater impacts for participants without recent work experience compared to those with recent work experience. However, the BCF approach suggested the opposite conclusion: there was a 63 percent chance that the impact was higher for those with recent work experience. Such differences arise because BCF estimates a completely different impact model from the standard approach (as discussed in Section II). The differing results signal that there is an additional level of uncertainty to the impact estimation, similar to when sensitivity analyses with different control variables do not match results from the main impact estimation. For this reason, we recommend putting more weight on subgroups that both BCF and the original impact estimation reveal as meaningful.

In most cases, the results from BCF did not suggest large differences between subgroups, potentially pointing to a lack of statistical power.

For most programs, the BCF approach identified the top subgroup variable as one for which there was little evidence of a difference in impacts between the two subgroups, either in terms of probabilities or the null hypothesis testing approach. For example, for FaDSS, BCF identified whether participants had recent employment as the best variable for defining subgroups with meaningful differences in impacts on self-reported monthly earnings. However, BCF also suggested that there was only a 60 percent chance that one of these two subgroups was larger than the other. Similarly,

based on the null hypothesis testing approach, the estimated difference in impacts was small ($70) between these subgroups and insignificant. Notably, for FaDSS, BCF did not identify other subgroups for which standard impact estimates differed by greater amounts. For example, the standard approach estimated a $265 difference in impacts by age for FaDSS that was statistically significant at the 10 percent level (Table 4). The inconsistency between the BCF approach and the estimated impact differences suggested that BCF may be identifying spurious subgroups, potentially due to a lack of statistical power. Consistent with this possibility, the one exception to this pattern was for MyGoals, the program with the largest sample size (nearly double that of the others). For MyGoals, BCF identified whether a participant was a single parent as the top variable on which to base subgroups. The estimated difference in impacts for this variable was $242, which was statistically significant. BCF also suggested a 75 percent chance that the impacts differed between these subgroups.

Table 5. Impacts of all four coaching programs on average monthly self-reported earnings for subgroups based on the BCF approach

					Probability that the true impact or difference in true impacts was...							
Program	Subgroup	Sample size	Control group mean	Estimated impact or difference and level of significance	Less than -$100	Less than -$50	Less than -$25	Less than $0	Greater than $0	Greater than $25	Greater than $50	Greater than $100
Employment status												
FaDSS	Not employed in the past month	162	892	24	0.00	0.03	0.07	0.19	0.81	0.64	0.50	0.24
	Employed in the past month	334	523	93	0.02	0.07	0.14	0.27	0.73	0.56	0.42	0.19
	Difference between subgroups			-70	0.01	0.05	0.11	0.40	0.60	0.27	0.17	0.07
Employment status												
Goal4 It!	Not employed in the past month	103	998	424*	0.04	0.11	0.17	0.28	0.72	0.59	0.48	0.30
	Employed in the past month	271	691	-41	0.02	0.06	0.11	0.20	0.80	0.68	0.59	0.42
	Difference between subgroups			465*	0.16	0.26	0.36	0.63	0.37	0.11	0.05	0.01
Language of the survey												
LIFT	Survey was not conducted in Spanish	321	752	49	0.12	0.30	0.43	0.63	0.37	0.20	0.11	0.03
	Survey was conducted in Spanish	253	991	-21	0.10	0.27	0.41	0.60	0.40	0.23	0.13	0.04
	Difference between subgroups			71	0.03	0.08	0.17	0.54	0.46	0.10	0.04	0.01
Single parent status												
MyGoals	Not a single parent	704	303	141***	0.00	0.01	0.04	0.17	0.83	0.60	0.40	0.13
	Single parent	544	481	-101	0.05	0.16	0.26	0.47	0.53	0.28	0.12	0.01
	Difference between subgroups			242***	0.00	0.00	0.00	0.25	0.75	0.42	0.32	0.19

Sources: Evaluation of Employment Coaching baseline and first follow-up surveys and MyGoals Baseline Questionnaire data.

Note: Outcomes are measured over the first 9 months after study enrollment for FaDSS, Goal4 It!, and LIFT. Outcomes are measured over the first 12 months after study enrollment for MyGoals. The estimated impacts and level of statistical significance come from the standard approach to estimating subgroup impacts based on the null hypothesis testing framework.

***/**/* Impact estimates are statistically significant at the .01/.05/.10 levels, respectively, using a two-tailed t-test.

Our simulation suggested that BCF requires relatively large sample sizes to identify meaningful differences in impacts between subgroups, especially when exploring many candidate subgroups.

When assuming a true difference in impacts for one subgroup and simulated data that approximated our analysis sample, we found that BCF infrequently identified a likely subgroup difference. In particular, BCF only identified a likely difference between subgroups 2 percent of the time, when using simulated data with 2,000 observations and 20 candidate subgroup variables (Table 6).[6] With a sample size of 5,000, the figure increased to 50 percent. With sample sizes of 10,000 and 20,000, BCF identified a likely difference between subgroups 93 percent and 100 percent of the time, respectively. In addition, with fewer candidate variables used to define subgroups, BCF identified likely subgroup differences at higher rates. These results were similar when we assumed a normally distributed outcome (Appendix Table B.8).

Table 6. Results from simulation of BCF for varying sample sizes and number of variables used to define subgroups

Number of candidate variables BCF used to define subgroups	Percentage of times BCF identified a likely subgroup difference (at least a 95 percent chance that the impacts differed between two subgroups)
Sample size of 2,000	
5	18
10	8
20	2
Sample size of 5,000	
5	80
10	55
20	50
Sample size of 10,000	
5	100
10	99
20	93
Sample size of 20,000	
5	100
10	100
20	100

Note: The estimates are based on (1) 100 simulation draws for each variable and sample size combination, (2) an assumed exponential distribution of earnings, and (3) a 0.20 effect size difference in impacts across one of the variables used to define the subgroups.

[6] Our sample sizes were smaller than the size assumed for this simulation and included more candidate subgroup variables, suggesting that our main analyses may have had even less statistical power.

DISCUSSION

In this section, we showed how the BCF approach can potentially identify subgroups that were not prespecified by evaluators. In our study of four employment coaching programs, we identified several new subgroups that had not been analyzed in the original evaluation. However, these subgroups did not appear to have meaningful impact differences according to the BCF analysis. In addition, with one exception, the subgroups that BCF identified did not correspond to statistically significant findings using the null hypothesis testing approach, even when we estimated large and statistically significant differences for other subgroups. These inconsistencies between BCF and standard impact estimates also highlighted why, unlike BHLM, BCF cannot be used to reinterpret impact estimates. BCF estimates a separate model, which can lead to discrepancies in findings arising from differing assumptions (with BCF having less stringent underlying assumptions) and model uncertainty, particularly in cases where the sample size is small. For this reason, we recommend using BCF to identify potentially new subgroups but confirming that they are meaningful in the original impact estimation framework.

These findings suggest that the Bayesian causal forest approach may have been underpowered for our application.

These findings suggest that the BCF approach may have been underpowered for our application. Consistent with that possibility, our simulations revealed that the BCF approach would identify likely subgroup differences only 2 percent of the time with data similar to what we used in this study. In addition, our simulations suggested that the power would decrease when more subgroups were examined. At the same time, the simulations revealed that BCF could identify likely subgroup differences at much higher rates with greater sample sizes—which suggests that it may be a promising approach for larger studies. This result was consistent with an earlier application of BCF that used a data set of over 6,000 individuals to identify new subgroups with differences in impacts (Hahn et al. 2020).

Our results suggest that evaluators could conduct a similar simulation to determine whether BCF would be likely to detect meaningful subgroup differences. In addition, with smaller sample sizes, evaluators may benefit from selecting a more limited set of candidate variables for BCF to consider when forming subgroups.

VII. Conclusions

This report illustrated several ways that evaluators can use Bayesian approaches to explore subgroup impacts of employment programs on outcomes of interest (Table 7). Our results showed how the BHLM approach could complement standard null hypothesis testing for a single evaluation by accounting for multiple comparisons, providing more nuanced conclusions, and guarding against small changes in the data or modeling decisions. The resulting probability statements could allow evaluators to provide some information even when results do not meet standard levels of statistical significance.

In addition, our analysis showed how BHLM could be used across multiple evaluations of employment programs or evaluations with multiple sites. When analyzing multiple programs simultaneously, our findings demonstrated how the BHLM method tends to shift the impact estimates toward each other while allowing for differences across programs or sites. This approach represents a middle ground—known as partial pooling—between treating programs or sites separately and completely pooling the data so that there is a single impact estimate across all programs or sites.

Whether to use partial pooling hinges on whether it is plausible that information about one program or site on subgroup impact estimates is relevant to other programs or sites. As with other analyses, we recommend deciding whether to use partial pooling across programs or sites prior to estimating impacts. When making this decision, evaluators may draw on additional information about the programs or sites, such as the similarity of the programs' logic models and services or program staff's experiences working with people from different subgroups. Evaluators pursuing this approach may also consider a sensitivity test that treats all programs or sites as entirely separate. Although the results of the sensitivity analysis would not inform whether the modeling assumptions were accurate, they would suggest the extent to which the assumptions made a practical difference.

Our exploration of the BCF approach suggested that the sample sizes for the Evaluation of Employment Coaching were too small for this approach to effectively identify subgroups that had not been prespecified. Additional simulations confirmed that our BCF analysis was likely underpowered but suggested that it may be promising with larger sample sizes. Evaluators considering the BCF approach may benefit from conducting similar simulations to determine the statistical power.

Table 7. Summary of analyses and considerations for evaluators of employment programs

Research question	Summary of results	Considerations for evaluators
How can Bayesian methods be used to reinterpret subgroup impact estimates for a single evaluation of an employment program?	• The BHLM approach can suggest more nuanced conclusions when individual subgroup estimates are not statistically significant, especially when accounting for multiple comparisons. • Conclusions based on the BHLM approach are less sensitive to small deviations in the data compared to conclusions based on the levels of statistical significance.	• BHLM can be a useful way to reinterpret subgroup impact estimates from a single evaluation. • The BHLM approach may be especially helpful for subgroup analyses with relatively low statistical power, either because they have small sample sizes or include many different hypothesis tests.
How can Bayesian methods be used to reinterpret subgroup impact estimates for multiple evaluations of employment programs?	• The impact estimates from separate programs influenced each other, highlighting how BHLM draws on information across programs. • The estimated probabilities still varied across programs, showcasing how BHLM can allow for different conclusions for different programs.	• Using the BHLM approach across multiple evaluations or sites can potentially increase statistical power. • Evaluators considering whether to use the BHLM approach across multiple programs or sites may wish to carefully consider the plausibility of using different programs or sites to inform each other.
To what extent can Bayesian methods be used in an evaluation of an employment program to identify subgroups that were not previously specified?	• The BCF approach can potentially identify different subgroups that had not been prespecified and allow evaluators to describe results in terms of probabilities. • In many cases, the results from BCF were inconsistent with the standard null hypothesis testing approach and did not suggest meaningful differences between subgroups, potentially pointing to a lack of statistical power. • Our simulation suggests that BCF requires relatively large sample sizes to identify meaningful differences in impacts between subgroups, especially when exploring many subgroups.	• BCF may be able to identify new subgroups, but it can also require a large sample size. • Evaluators may consider conducting a simulation to determine whether BCF would be likely to detect meaningful subgroup differences. • Evaluators may benefit from selecting a limited set of candidate subgroup variables for BCF to consider, especially if sample sizes are small.

Three limitations of this study suggest directions for future research and the application of these methods for subgroup analysis:

1. Our analysis used data from a single study of four different employment programs, so the performance of the Bayesian methods we studied may differ when using other data. For example, our exploration of the BCF approach did not yield strong evidence of differences in subgroup estimates. However, our simulation study suggested that our sample size may have been too small to use BCF effectively. Other evaluations with larger subgroup differences or larger sample sizes may benefit more from using BCF. Future work could explore the performance of these methods using different data sources.
2. The simulations that we conducted to understand the statistical power of the BCF approach focused on a limited set of scenarios that were most germane to our analyses. Thus, the findings may not apply in other situations. For example, our target effect size of 0.20 may differ in other applications. Evaluators considering this approach could conduct similar calculations before pursuing BCF to determine whether it would have sufficient statistical power.
3. This report focused on two of many possible methods for extending subgroup analyses beyond the standard null hypothesis testing approach. We selected these methods because they have been shown to have some advantages over other approaches. However, there are alternative approaches that evaluators could consider. For example, for evaluations with relatively few subgroups—and therefore less potential for the multiple comparisons problem—evaluators could use Bayesian methods to estimate impacts for each subgroup separately. Similarly, there are a number of other methods aside from BCF that can flexibly estimate impacts with few parametric assumptions, including targeted maximum likelihood estimation (Schuler and Rose 2017), gradient boosting (Friedman 2001), and X learner (Künzel et al. 2018). Future research could explore some of these approaches to provide evaluators with more tools that generate nuanced information about subgroup impact estimates and better inform policymakers and practitioners about how to serve and design programs for specific groups.

References

Benjamini, Y., and Y. Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57 no. 1, 1995, pp. 289–300.

Breiman, L., J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees*. New York: Routledge, 1984.

Deke, J., and M. Finucane. "Moving Beyond Statistical Significance: The BASIE (BAyeSian Interpretation of Estimates) Framework for Interpreting Findings from Impact Evaluations." OPRE Report #2019-35. Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services, 2019.

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

Dorie, V., J. Hill, U. Shalit, M. Scott, and D. Cervone. "Automated Versus Do-It-Yourself Methods for Causal Inference: Lessons Learned from a Data Analysis Competition." *Statistical Science*, vol. 34, no. 1, 2019, pp. 43–68.

Friedman, J. H. "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics*, vol. 29, no. 5, 2001, pp. 1189–1232.

Gelman, A. "P Values and Statistical Practice." *Epidemiology*, vol. 24, no. 1, 2013, pp. 69–72.

Gelman, A., and H. Stern. "The Difference Between 'Significant' and 'Not Significant' Is Not Itself Statistically Significant." *The American Statistician*, vol. 60, no. 4, 2006, pp. 328–331.

Gelman, A., J. Hill, and M. Yajima. "Why We (Usually) Don't Have to Worry About Multiple Comparisons." *Journal of Research on Educational Effectiveness*, vol. 5, no. 2, 2012, pp. 965–1056.

Gigerenzer, G. "Statistical Rituals: The Replication Delusion and How We Got There." *Advances in Methods and Practices in Psychological Science*, vol. 1, no. 2, 2018, pp. 198–218.

Goodman, Steven N. "Aligning Statistical and Scientific Reasoning." *Science*, vol. 352, no. 6290, 2016, pp. 1180–1181.

Greenland, S., S.J. Senn, K.J. Rothman, J.B. Carlin, C. Poole, S.N. Goodman, and D.G. Altman. "Statistical Tests, *P*-Values, Confidence Intervals, and Power: A Guide to Misinterpretations." *European Journal of Epidemiology*, vol. 31, no. 4, 2016, pp. 337–350.

Hahn, P.R., J.S. Murray, and C.M. Carvalho. "Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects (with Discussion)." *Bayesian Analysis*, vol. 15, no. 3, 2020, pp. 965–1056.

Hill, J. "Bayesian Nonparametric Modeling for Causal Inference," *Journal of Computational and Graphical Statistics*, vol. 20, no. 1, 2011, pp. 217-240.

Joyce, K., and McConnell, S. "Employment Coaching for TANF and Related Populations." OPRE Report #2019-67. Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services, 2019.

Künzel, S. R., J. S. Sekhon, P. J. Bickel, and B. Yu. "Metalearners for Estimating Heterogenous Treatment Effects Using Machine Learning." *Proceedings of the National Academy of Sciences*, vol. 116, no. 10, 2018, pp. 4156–4165.

Lipman, E.R., J. Deke, and M.M. Finucane. "Bayesian Interpretation of Cluster Robust Subgroup Impact Estimates: The Best of Both Worlds." *Journal of Policy Analysis and Management*, vol. 41, no. 4, 2022, pp. 1204–1224.

Moore, Q., T. Kautz, S. McConnell, O. Schochet, and A. Wu. "Can a Participant-Centered Approach to Setting and Pursuing Goals Help Adults with Low Incomes Become Economically Stable? Short-Term Impacts of Four Employment Coaching Programs." OPRE Report #2023-139. Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services, 2023.

Porter, K. E. "Statistical Power in Evaluations That Investigate Effects on Multiple Outcomes: A Guide for Researchers." *Journal of Research on Educational Effectiveness*, vol. 11, no. 2, 2018, pp. 267–295.

Schuler, M. S., and S. Rose. "Targeted Maximum Likelihood Estimation for Causal Inference in Observational Studies." *American Journal of Epidemiology*, vol. 185, no. 1, 2017, pp. 65–73.

Shiferaw, L., and D. Thal. "Digging Deeper into What Works: What Services Improve Labor Market Outcomes, and for Whom?" OPRE Report #2022-161. Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services, 2022.

Thal, D., and M. Finucane. "Causal Methods Madness: Lessons Learned from the 2022 ACIC Competition to Estimate Health Policy Impacts." *Observational Studies*, vol. 9, no. 3, 2023, pp. 3–27.

Tukey, J.W. "The Problem of Multiple Comparisons." In *The Collected Works of John W. Tukey VIII, Multiple Comparisons: 1948—1983*, edited by H.I. Braun. New York: Chapman and Hall, May 1953.

Vollmer, L., M. Finucane, and R. Brown. "Revolutionizing Estimation and Inference for Program Evaluation Using Bayesian Methods." *Evaluation Review*, vol. 44, no. 4, 2020, pp. 295–324.

Wasserstein, R.L., and N.A. Lazar. "The ASA's Statement on *P*-Values: Context, Process, and Purpose." *The American Statistician*, vol. 70, no. 2, 2016, pp. 129–133.

Zellner, A. "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias." *Journal of the American Statistical Association*, vol. 57, no. 298, 1962, pp. 348–368.

Appendix A: Technical Notes

STANDARD SUBGROUP IMPACT ESTIMATION METHODS BASED ON NULL HYPOTHESIS TESTING

Following Moore (et al. 2023), we estimated subgroup impacts from a joint model that pooled data for the two subgroups and included indicators for the research group and the subgroups as well as an interaction between the research group indicator and the subgroup indicator. We used ordinary least squares to estimate the following linear regression:

$$y_i = \alpha + \beta_1 subgroup_i + \beta_2 program_i + \beta_3 program_i subgroup_i + \gamma X_i + \varepsilon_i,$$

where y_i is the outcome of interest for participant i, $subgroup_i$ is an indicator for whether a participant was in the subgroup or not, $program_i$ is an indicator for whether a participant was assigned to the program group or the control group, X_i is a vector of baseline covariates, and ε_i is a random error term. In this specification, β_2 represents the impact of the program for someone who is not in $subgroup_i$ and $\beta_2 + \beta_3$ represents the impact for someone who is in $subgroup_i$. The baseline covariates included (1) measures of self-reported earnings, (2) characteristics that were observed to differ between the program and control group members, (3) indicators for when participants enrolled in the study, and (4) indicators of program location (when applicable). See Moore et al. (2023) for additional details on these covariates. For our reporting of standard impact estimates, we used the sign and statistical significance of the interaction (β_3).

BAYESIAN HIERARCHICAL LINEAR MODEL

Estimating BHLM is a three-step process. In the first step, we estimated standard linear models to obtain point estimates of the treatment impacts for each subgroup in each evaluation. In the second step, we used seemingly unrelated estimation on the models from the first step to estimate the error covariance between all of our estimates. In the third step, we used all of the estimates from the first step and their covariance from the second step to estimate the titular BHLM, which returned the posterior probabilities that reflected the partial pooling of the BHLM.

Step 1: Standard subgroup estimation based on null hypothesis testing

Following the approach for standard subgroup estimation based on null hypothesis testing described above, we estimated standard subgroup impacts and standard errors.

Step 2: Seemingly unrelated estimation

To estimate the covariance between impact estimates, we used seemingly unrelated estimation, as proposed by Zellner (1962) (and as implemented by the suest command in Stata 17). This method combined the parameters and residuals from all of the separate estimated models to estimate the covariance between parameters from different models and allowed us to calculate the error covariance of the different subgroup treatment estimates.

Because the models were run entirely separately for each of the four evaluations, they had no shared sample and no shared coefficients. So, by definition, there was no error covariance between the cross-evaluation impact estimates. Therefore, we ran the seemingly unrelated estimation separately for each of the four evaluations and combined them into a block-diagonal matrix if we needed a covariance matrix for all evaluations.

Step 3: Estimating BHLM

We assumed the following form for the BHLM:

$$y_{ij} \sim MVN\left(\alpha + \beta X_i + \delta_j + \phi_{ij}, \Sigma\right),$$

where y_{ij} are all of the impact estimates, on the effect size scale, indexed with subgroup i and evaluation j and Σ is the covariance matrix of those estimates (from Step 2). This model has the flavor of a meta-analytic model with a grand mean impact α common to all evaluations, covariate-explainable impacts β (where X_i is a matrix of subgroup compositions for each estimate), evaluation effects δ_j, and idiosyncratic subgroup-by-evaluation effects ϕ_{ij}.

Because our estimates were for partially overlapping subgroups and we estimated conditional impacts rather than marginal ones (that is, we did not hold other subgroup variables constant when estimating the impact in a given subgroup), we had to account for the fact that there was not only error covariance between estimates but also covariance in true impacts. For example, consider that age and education are correlated, with more educated workers tending to be older. If we found that there were larger effects for older workers, then this age effect would cause the subgroup estimate for more educated workers to be larger as well. Therefore, to capture the differential age distribution across levels of education, we created the matrix X_i. More broadly, X_i reflected the composition of each subgroup in terms of all of the subgroups (for example, the gender, age, education, disability, and so on compositions for each of the gender, age, education, disability, and so on subgroups), which was used by the model to estimate the marginal effects, β, for each subgroup.

Because this is a Bayesian model, evaluators need to carefully choose a prior for each parameter in the model. For the main parameters of interest ($\alpha, \beta, \delta, \varepsilon$), evaluators can specify a prior distribution in terms of other parameters called hyperparameters. For example, our prior for β was that it was normally distributed with a standard deviation, σ_β. This standard deviation was a hyperparameter, a parameter that governs the distribution of another parameter. We relied heavily on a Bayesian meta-analysis of the Pathways to Work Evidence Clearinghouse (Shiferaw and Thal 2022) to inform our priors and chose hyperparameter values. This meta-analysis estimated average effects across all interventions, the variance of estimate-specific effects, and the variance of combined evaluation and intervention effects as well as various effects for the population served, services offered, and their interactions (Appendix Tables A.1 and A.2). Because the model operates on standardized estimates, all priors are specified in terms of standard deviation units.

Table A.1. Hierarchical priors used for the BHLM analysis

Parameter	Description	Prior distribution
α	Overall effect of coaching interventions	$SGT(\mu_\alpha,\sigma_\alpha,\lambda_\alpha,\nu_\alpha,2)$
β	Marginal subgroup effects	$N(0,\sigma_\beta)$
δ	Evaluation random effects	$SGT(0,\sigma_\delta,\lambda_\delta,\nu_\delta,2)$
ε	Estimate random effects	$SGT(0,\sigma_\varepsilon,\lambda_\varepsilon,\nu_\varepsilon,2)$

Note: $N(\mu,\sigma)$ indicates a normal distribution with mean μ and standard deviation σ. $SGT(\mu,\sigma,\lambda,\nu,p)$ indicates a skewed generalized t distribution with location μ, scale σ, skewness λ, degrees of freedom ν, and kurtosis parameter p.

Table A.2. Hyperparameter estimates from Shiferaw and Thal (2022) used for the BHLM analysis

Parameter	Description	Hyperparameter value	Analogous parameter (Shiferaw and Thal 2022)
μ_α	Mean of overall effect	0.015	θ^{Int}
σ_α	Standard deviation of overall effect	0.018	$\frac{\sigma^{Study}}{\sqrt{2}}$
λ_α	Skewness of overall effect	0.37	λ^{Study}
ν_α	Degrees of freedom of overall effect	11.8	ν^{Study}
σ_β	Standard deviation of subgroup effects	0.02	$\sqrt{\sigma^{Foc\,2}+\sigma^{Foc*Foc\,2}+\sigma^{PrimFoc\,2}}$
σ_δ	Standard deviation of evaluation effects	0.018	$\frac{\sigma^{Study}}{\sqrt{2}}$
λ_δ	Skewness of evaluation effects	0.37	λ^{Study}
ν_δ	Degrees of freedom of evaluation effects	11.8	ν^{Study}
σ_ε	Standard deviation of estimate effects	0.025	σ^{Find}
λ_ε	Skewness of estimate effects	0.51	λ^{Find}
ν_ε	Degrees of freedom of estimate effects	3.1	ν^{Find}

Note: Because Shiferaw and Thal (2022) only estimated a combined evaluation plus intervention effect while we estimated separate ones, we apportioned the combined variance (σ^{Study^2}) estimated by Shiferaw and Thal (2022) equally between σ_α^2 and σ_δ^2. Because Shiferaw and Thal (2022) estimated both focus population main effects, interactions between multiple focus populations, and interactions between focus population and primary services offered, we used the combined standard deviation across all three of those batches of effects to inform the standard deviation for our subgroup effects.

The BHLM equation outlined above allowed for both evaluation and subgroup effects. To estimate the model for subgroups within a single evaluation, we simply fit the same model with a single evaluation effect shared by all estimates. Given the priors, this was mathematically equivalent to omitting the δ_j term and using a prior on α of $SGT\left(\mu_\alpha, \sqrt{\sigma_\alpha^2 + \sigma_\delta^2}, \lambda_\alpha, \nu_\alpha, 2\right)$.

BAYESIAN CAUSAL FOREST

BCF (Hahn et al. 2020) is an extension of the Bayesian Additive Regression Tree (BART) model introduced by Hill (2011). BART is a fully Bayesian regression tree model fit via Markov chain Monte Carlo (MCMC). The fact that it is a nonparametric regression tree allows for tremendous flexibility in identifying the functional form of relationships between covariates and outcomes. BART pairs the regression tree approach with Bayesian priors on both the tree structure (with a prior that prefers simpler, shallower trees) and the predicted values for each leaf node (employing Bayesian partial pooling to shift predictions for nodes back to the grand mean in inverse proportion to how strong the signal in each node is).

BCF extends BART specifically to causal inference by simultaneously fitting two BART models: one, called μ below, to flexibly identify relationships between covariates (in our case, individual characteristics X_i and indicators for which evaluation the participant was part of I_j) and the outcome; and another, called τ below, to flexibly identify the relationship between covariates and treatment impacts:

$$y_{ij} = \mu(X_i, I_j) + z_i \tau(X_i, I_j) + \varepsilon_i .$$

Typically, BCF also includes an estimated propensity score in the μ function as a single-variable summary of covariate-treatment confounding relationships. Because this was an experimental evaluation, however, we omitted the propensity score from the model (it was constant across all individuals).

BCF produces estimated treatment impact for each individual, τ_i. We estimated the impact in each prespecified subgroup S by simply averaging the individual-level impacts across relevant individuals:

$$\tau_s = \frac{1}{n_s} \sum_{i \in S} \tau_i .$$

However, because BCF produces these individual-level estimates, it can go much further than this and allow for discovery of non-prespecified subgroups. To do this, we took a fit-the-fit approach, where first we estimated the model and calculated each individual's treatment estimate, τ_i. Next, we fit a CART model to those estimates to identify different groupings of individuals with high treatment impact estimates. Lastly, we estimated subgroup impacts for each of those groupings by applying the same calculation outlined above but using our CART-constructed indicator in place of the prespecified subgroup S.

Appendix B: Supplementary Information

Table B.1. Impacts of a single coaching program (Goal4 It!) on average monthly self-reported earnings by subgroup based on the BHLM approach

				Probability that the true impact or difference in true impacts was...							
Subgroup	Sample size	Control group mean	Estimated impact or difference and level of significance	Less than -$100	Less than -$50	Less than -$25	Less than $0	Greater than $0	Greater than $25	Greater than $50	Greater than $100
Participant age											
Age 30 or younger	188	452	351***	0.00	0.00	0.02	0.13	0.87	0.65	0.38	0.07
Older than age 30	251	983	-146	0.00	0.01	0.07	0.30	0.70	0.36	0.13	0.01
Difference between subgroups			497***	0.00	0.01	0.07	0.26	0.74	0.47	0.23	0.04
Number of children											
Fewer than two	194	504	347**	0.00	0.00	0.04	0.20	0.80	0.52	0.24	0.03
Two or more	242	941	-131	0.00	0.01	0.07	0.30	0.70	0.38	0.14	0.01
Difference between subgroups			478**	0.00	0.02	0.11	0.35	0.65	0.31	0.10	0.01
Education level											
No college	216	478	193*	0.00	0.00	0.03	0.17	0.83	0.55	0.28	0.03
Some college or higher	223	964	61	0.00	0.00	0.04	0.21	0.79	0.50	0.24	0.03
Difference between subgroups			132	0.01	0.07	0.20	0.45	0.55	0.27	0.10	0.01
Race and ethnicity											
Not Hispanic/ Latino/a	208	714	146	0.00	0.00	0.04	0.19	0.81	0.55	0.29	0.05
Hispanic/ Latino/a	145	737	109	0.00	0.00	0.04	0.18	0.82	0.56	0.31	0.08
Difference between subgroups			37	0.04	0.13	0.27	0.52	0.48	0.24	0.10	0.02
Goal-setting skills											
Above median score	259	984	35	0.00	0.01	0.07	0.32	0.68	0.34	0.12	0.01
At or below median score	164	448	16	0.00	0.02	0.10	0.34	0.66	0.34	0.13	0.01
Difference between subgroups			19	0.00	0.03	0.17	0.49	0.51	0.18	0.04	0.00

(continued)

Subgroup	Sample size	Control group mean	Estimated impact or difference and level of significance	Probability that the true impact or difference in true impacts was... Less than -$100	Less than -$50	Less than -$25	Less than $0	Greater than $0	Greater than $25	Greater than $50	Greater than $100
Barriers to employment											
At or below median score	182	850	147	0.00	0.01	0.07	0.28	0.72	0.41	0.17	0.01
Above median score	220	663	31	0.00	0.01	0.07	0.31	0.69	0.37	0.14	0.01
Difference between subgroups			116	0.00	0.04	0.18	0.47	0.54	0.23	0.06	0.00
Extent to which transportation made it hard to find employment											
A little, not at all	162	749	337**	0.00	0.00	0.04	0.18	0.82	0.56	0.29	0.04
Somewhat, very, or extremely	238	770	-112	0.00	0.01	0.08	0.33	0.67	0.34	0.12	0.01
Difference between subgroups			489**	0.00	0.02	0.10	0.31	0.69	0.39	0.16	0.02
Extent to which lack of childcare made it hard to find employment											
A little, not at all	151	838	156	0.00	0.00	0.03	0.15	0.85	0.62	0.38	0.09
Somewhat, very, or extremely	248	663	114	0.00	0.00	0.03	0.17	0.83	0.57	0.30	0.04
Difference between subgroups			42	0.01	0.10	0.24	0.45	0.55	0.33	0.17	0.04
Driver's license											
Has driver's license	138	488	166	0.00	0.01	0.06	0.25	0.75	0.45	0.21	0.02
Does not have driver's license	252	939	22	0.00	0.01	0.06	0.29	0.71	0.39	0.15	0.01
Difference between subgroups			144	0.00	0.03	0.16	0.44	0.56	0.25	0.08	0.01
Employment status											
Employed in the past month	103	998	424*	0.00	0.01	0.06	0.24	0.76	0.47	0.21	0.03
Not employed in the past month	271	691	-41	0.00	0.01	0.07	0.31	0.69	0.36	0.13	0.01
Difference between subgroups			465*	0.00	0.02	0.12	0.41	0.60	0.27	0.09	0.01

Sources: Evaluation of Employment Coaching baseline and first follow-up surveys.

Note: Outcomes are measured over the first 9 months after study enrollment. Estimated impacts and level of statistical significance come from the standard approach to estimating subgroup impacts based on the null hypothesis testing framework.

***/**/* Impact estimates are statistically significant at the .01/.05/.10 levels, respectively, using a two-tailed t-test.

Table B.2. Impacts of a single coaching program (FaDSS) on average monthly self-reported earnings by subgroup based on the BHLM approach

				Probability that the true impact or difference in true impacts was...							
Subgroup	Sample size	Control group mean	Estimated impact or difference and level of significance	Less than -$100	Less than -$50	Less than -$25	Less than $0	Greater than $0	Greater than $25	Greater than $50	Greater than $100
Participant age											
Age 30 or younger	289	559	199**	0.00	0.00	0.02	0.16	0.84	0.53	0.22	0.01
Older than age 30	208	737	-66	0.00	0.01	0.08	0.32	0.68	0.36	0.13	0.01
Difference between subgroups			265*	0.00	0.02	0.11	0.32	0.68	0.35	0.11	0.00
Number of children											
Two or more	321	665	93	0.00	0.00	0.07	0.33	0.67	0.31	0.09	0.00
Fewer than two	176	613	13	0.00	0.01	0.09	0.35	0.65	0.33	0.11	0.01
Difference between subgroups			80	0.00	0.03	0.17	0.51	0.49	0.15	0.02	0.00
Education level											
Some college or higher	179	663	96	0.00	0.01	0.07	0.27	0.73	0.41	0.16	0.01
No college	318	635	54	0.00	0.00	0.06	0.30	0.70	0.33	0.10	0.00
Difference between subgroups			42	0.00	0.02	0.13	0.44	0.56	0.22	0.05	0.00
Race and ethnicity											
White	256	675	104	0.00	0.00	0.04	0.21	0.79	0.49	0.22	0.02
Not white	234	618	58	0.00	0.00	0.04	0.22	0.78	0.46	0.19	0.01
Difference between subgroups			46	0.00	0.05	0.19	0.47	0.53	0.24	0.07	0.00
Goal-setting skills											
Above median score	298	703	105	0.00	0.00	0.03	0.20	0.80	0.47	0.18	0.01
At or below median score	195	524	87	0.00	0.00	0.05	0.23	0.77	0.47	0.20	0.02
Difference between subgroups			18	0.00	0.06	0.21	0.49	0.51	0.20	0.05	0.00
Barriers to employment											
Above median score	208	495	120	0.00	0.01	0.06	0.25	0.75	0.44	0.17	0.01
At or below median score	289	746	40	0.00	0.00	0.06	0.27	0.73	0.39	0.13	0.01
Difference between subgroups			80	0.00	0.04	0.17	0.46	0.54	0.24	0.07	0.00

(continued)

Subgroup	Sample size	Control group mean	Estimated impact or difference and level of significance	Probability that the true impact or difference in true impacts was... Less than -$100	Less than -$50	Less than -$25	Less than $0	Greater than $0	Greater than $25	Greater than $50	Greater than $100
Extent to which transportation made it hard to find employment											
Somewhat, very, or extremely	255	552	127	0.00	0.01	0.06	0.29	0.71	0.37	0.13	0.00
A little, not at all	240	745	-1	0.00	0.01	0.09	0.34	0.66	0.32	0.11	0.00
Difference between subgroups			128	0.00	0.03	0.15	0.44	0.56	0.23	0.05	0.00
Extent to which lack of childcare made it hard to find employment											
Somewhat, very, or extremely	258	533	177*	0.00	0.00	0.03	0.16	0.84	0.55	0.25	0.02
A little, not at all	238	746	-25	0.00	0.01	0.07	0.28	0.72	0.39	0.15	0.01
Difference between subgroups			202	0.00	0.03	0.13	0.26	0.64	0.35	0.12	0.00
Driver's license											
Has driver's license	215	464	104	0.00	0.00	0.06	0.27	0.73	0.40	0.15	0.01
Does not have driver's license	282	798	19	0.00	0.01	0.09	0.36	0.64	0.30	0.09	0.00
Difference between subgroups			84	0.00	0.02	0.11	0.39	0.61	0.26	0.06	0.00
Employment status											
Not employed in the past month	334	523	93	0.00	0.00	0.06	0.31	0.69	0.32	0.10	0.00
Employed in the past month	162	892	24	0.00	0.01	0.09	0.32	0.68	0.35	0.13	0.01
Difference between subgroups			70	0.00	0.03	0.16	0.51	0.49	0.15	0.02	0.00
Urbanicity											
Rural	148	640	190	0.00	0.01	0.06	0.25	0.75	0.45	0.19	0.02
Urban	349	631	47	0.00	0.00	0.05	0.28	0.72	0.36	0.11	0.00
Difference between subgroups			142	0.00	0.02	0.14	0.43	0.57	0.25	0.07	0.00

Sources: Evaluation of Employment Coaching baseline and first follow-up surveys.

Note: Outcomes are measured over the first 9 months after study enrollment. Estimated impacts and level of statistical significance come from the standard approach to estimating subgroup impacts based on the null hypothesis testing framework.

***/**/* Impact estimates are statistically significant at the .01/.05/.10 levels, respectively, using a two-tailed t-test.

Table B.3. Impacts of a single coaching program (LIFT) on average monthly self-reported earnings by subgroup based on the BHLM approach

				Probability that the true impact or difference in true impacts was...							
Subgroup	Sample size	Control group mean	Estimated impact or difference and level of significance	Less than -$100	Less than -$50	Less than -$25	Less than $0	Greater than $0	Greater than $25	Greater than $50	Greater than $100
Participant age											
Age 30 or younger	198	790	36	0.00	0.01	0.09	0.33	0.67	0.33	0.11	0.00
Older than age 30	376	901	13	0.00	0.00	0.07	0.34	0.66	0.30	0.08	0.00
Difference between subgroups			23	0.00	0.04	0.18	0.49	0.51	0.20	0.05	0.00
Number of children											
Two or more	419	727	30	0.00	0.00	0.06	0.34	0.66	0.30	0.08	0.00
Fewer than two	155	1,207	-7	0.00	0.02	0.11	0.35	0.65	0.35	0.13	0.01
Difference between subgroups			37	0.00	0.06	0.21	0.51	0.49	0.19	0.05	0.00
Education level											
No college	372	593	38	0.00	0.00	0.07	0.33	0.67	0.32	0.09	0.00
Some college or higher	202	1,326	0	0.00	0.02	0.11	0.35	0.65	0.35	0.14	0.01
Difference between subgroups			38	0.00	0.07	0.23	0.50	0.50	0.22	0.06	0.00
Race and ethnicity											
Hispanic/ Latino/a	419	568	65	0.00	0.00	0.04	0.22	0.78	0.46	0.19	0.01
Not Hispanic/ Latino/a	153	1,535	35	0.00	0.03	0.12	0.33	0.67	0.42	0.20	0.03
Difference between subgroups			30	0.02	0.10	0.23	0.44	0.56	0.32	0.14	0.01
Goal-setting skills											
Above median score	346	907	28	0.00	0.00	0.06	0.31	0.69	0.33	0.09	0.00
At or below median score	221	789	19	0.00	0.01	0.08	0.32	0.68	0.35	0.12	0.01
Difference between subgroups			9	0.00	0.05	0.20	0.50	0.50	0.19	0.04	0.00
Barriers to employment											
Above median score	260	687	53	0.00	0.00	0.06	0.26	0.74	0.43	0.18	0.01
At or below median score	308	1,003	22	0.00	0.00	0.05	0.24	0.76	0.42	0.16	0.01
Difference between subgroups			30	0.00	0.08	0.24	0.51	0.49	0.24	0.08	0.00

(continued)

Subgroup	Sample size	Control group mean	Estimated impact or difference and level of significance	Probability that the true impact or difference in true impacts was... Less than -$100	Less than -$50	Less than -$25	Less than $0	Greater than $0	Greater than $25	Greater than $50	Greater than $100
Extent to which transportation made it hard to find employment											
Somewhat, very, or extremely	224	683	25	0.00	0.01	0.11	0.37	0.63	0.29	0.09	0.00
A little, not at all	347	993	-1	0.00	0.01	0.12	0.43	0.57	0.23	0.06	0.00
Difference between subgroups			26	0.00	0.03	0.15	0.45	0.55	0.23	0.06	0.00
Extent to which lack of childcare made it hard to find employment											
Somewhat, very, or extremely	347	719	92	0.00	0.00	0.06	0.26	0.74	0.41	0.15	0.01
A little, not at all	223	1,075	-50	0.00	0.01	0.09	0.31	0.69	0.37	0.14	0.01
Difference between subgroups			142	0.00	0.06	0.20	0.46	0.54	0.26	0.08	0.00
Driver's license											
Does not have driver's license	226	1,250	43	0.00	0.01	0.10	0.34	0.66	0.35	0.13	0.01
Has driver's license	348	584	-8	0.00	0.01	0.08	0.36	0.64	0.28	0.08	0.00
Difference between subgroups			51	0.00	0.04	0.17	0.46	0.54	0.23	0.07	0.00
Employment status											
Not employed in the past month	293	226	43	0.00	0.00	0.05	0.27	0.73	0.39	0.14	0.01
Employed in the past month	280	1,470	39	0.00	0.01	0.09	0.31	0.69	0.38	0.15	0.01
Difference between subgroups			4	0.01	0.08	0.21	0.47	0.53	0.24	0.08	0.00
Preferred language											
Spanish	321	447	72	0.00	0.01	0.07	0.30	0.70	0.35	0.12	0.00
English	253	1,350	-54	0.00	0.02	0.13	0.40	0.60	0.29	0.09	0.00
Difference between subgroups			127	0.00	0.05	0.18	0.42	0.58	0.30	0.11	0.00

(continued)

Subgroup	Sample size	Control group mean	Estimated impact or difference and level of significance	Probability that the true impact or difference in true impacts was... Less than -$100	Less than -$50	Less than -$25	Less than $0	Greater than $0	Greater than $25	Greater than $50	Greater than $100
LIFT program location											
Los Angeles	332	691	26	0.00	0.01	0.08	0.37	0.63	0.27	0.07	0.00
Chicago or New York	242	1,082	17	0.00	0.02	0.12	0.39	0.61	0.29	0.09	0.00
Difference between subgroups			9	0.00	0.05	0.20	0.49	0.51	0.22	0.06	0.00

Sources: Evaluation of Employment Coaching baseline and first follow-up surveys.

Note: Outcomes are measured over the first 9 months after study enrollment. Estimated impacts and level of statistical significance come from the standard approach to estimating subgroup impacts based on the null hypothesis testing framework.

***/**/* Impact estimates are statistically significant at the .01/.05/.10 levels, respectively, using a two-tailed t-test.

Table B.4. Impacts of a single coaching program (MyGoals) on average monthly self-reported earnings by subgroup based on the BHLM approach

				Probability that the true impact or difference in true impacts was...							
Subgroup	Sample size	Control group mean	Estimated impact or difference and level of significance	Less than -$100	Less than -$50	Less than -$25	Less than $0	Greater than $0	Greater than $25	Greater than $50	Greater than $100
Participant age											
Older than age 30	962	386	56	0.00	0.00	0.02	0.22	0.78	0.34	0.07	0.00
Age 30 or younger	289	357	-34	0.00	0.01	0.09	0.34	0.66	0.32	0.10	0.00
Difference between subgroups			90	0.00	0.03	0.14	0.42	0.58	0.22	0.04	0.00
Number of children											
Fewer than two	637	298	79	0.00	0.00	0.03	0.21	0.79	0.41	0.11	0.00
Two or more	583	472	-21	0.00	0.01	0.09	0.38	0.62	0.25	0.05	0.00
Difference between subgroups			100	0.00	0.01	0.08	0.33	0.67	0.31	0.08	0.00
Education level											
Some college or higher	493	433	72	0.00	0.00	0.04	0.24	0.76	0.40	0.12	0.00
No college	758	345	6	0.00	0.00	0.06	0.33	0.67	0.26	0.05	0.00
Difference between subgroups			66	0.00	0.01	0.10	0.37	0.63	0.26	0.06	0.00
Employment status											
Not employed in the past month	679	211	38	0.00	0.00	0.03	0.24	0.76	0.36	0.08	0.00
Employed in the past month	568	568	33	0.00	0.00	0.05	0.27	0.73	0.37	0.11	0.00
Difference between subgroups			5	0.00	0.04	0.18	0.50	0.50	0.18	0.03	0.00
Disability status											
Does not have disability	879	446	29	0.00	0.00	0.06	0.38	0.62	0.21	0.04	0.00
Has disability	344	221	4	0.00	0.01	0.09	0.35	0.65	0.29	0.07	0.00
Difference between subgroups			25	0.00	0.03	0.19	0.54	0.46	0.13	0.02	0.00
MyGoals program location											
Houston	724	423	38	0.00	0.00	0.05	0.29	0.71	0.32	0.07	0.00
Baltimore	528	319	26	0.00	0.00	0.05	0.26	0.74	0.36	0.10	0.00
Difference between subgroups			12	0.00	0.04	0.19	0.53	0.47	0.15	0.03	0.00

Sources: Evaluation of Employment Coaching first follow-up survey and MyGoals Baseline Questionnaire data.

Note: Outcomes are measured over the first 12 months after study enrollment. Estimated impacts and level of statistical significance come from the standard approach to estimating subgroup impacts based on the null hypothesis testing framework.

***/**/* Impact estimates are statistically significant at the .01/.05/.10 levels, respectively, using a two-tailed t-test.

Table B.5. Impacts of all four coaching programs on average monthly self-reported earnings by employment status at baseline based on the BHLM approach

					Probability that the true impact or difference in true impacts was...							
Program	Subgroup	Sample size	Control group mean	Estimated impact or difference and level of significance	Less than -$100	Less than -$50	Less than -$25	Less than $0	Greater than $0	Greater than $25	Greater than $50	Greater than $100
(a) Estimating separately for each program												
FaDSS	Not employed in the past month	334	523	93	0.00	0.00	0.06	0.31	0.69	0.32	0.10	0.00
	Employed in the past month	162	892	24	0.00	0.01	0.09	0.32	0.68	0.35	0.13	0.01
	Difference between subgroups			70	0.00	0.03	0.16	0.51	0.49	0.15	0.02	0.00
Goal4 It!	Not employed in the past month	271	691	-41	0.00	0.01	0.07	0.31	0.69	0.36	0.13	0.01
	Employed in the past month	103	998	424*	0.00	0.01	0.06	0.24	0.76	0.47	0.21	0.03
	Difference between subgroups			-465*	0.01	0.09	0.27	0.60	0.41	0.12	0.02	0.00
LIFT	Not employed in the past month	293	226	43	0.00	0.00	0.05	0.27	0.73	0.39	0.14	0.01
	Employed in the past month	280	1,470	39	0.00	0.01	0.09	0.31	0.69	0.38	0.15	0.01
	Difference between subgroups			4	0.01	0.08	0.21	0.47	0.53	0.24	0.08	0.00
MyGoals	Not employed in the past month	679	211	38	0.00	0.00	0.03	0.24	0.76	0.36	0.08	0.00
	Employed in the past month	568	568	33	0.00	0.00	0.05	0.27	0.73	0.37	0.11	0.00
	Difference between subgroups			5	0.00	0.04	0.18	0.50	0.50	0.18	0.03	0.00

(continued)

Program	Subgroup	Sample size	Control group mean	Estimated impact or difference and level of significance	Probability that the true impact or difference in true impacts was... Less than -$100	Less than -$50	Less than -$25	Less than $0	Greater than $0	Greater than $25	Greater than $50	Greater than $100
(b) Combining information across programs												
FaDSS	Not employed in the past month	334	523	93	0.00	0.00	0.03	0.21	0.79	0.41	0.12	0.00
	Employed in the past month	162	892	24	0.00	0.01	0.05	0.23	0.77	0.45	0.19	0.01
	Difference between subgroups			70	0.00	0.04	0.18	0.52	0.48	0.14	0.02	0.00
Goal4 It!	Not employed in the past month	271	691	-41	0.00	0.00	0.03	0.21	0.79	0.45	0.16	0.01
	Employed in the past month	103	998	424*	0.00	0.00	0.04	0.19	0.81	0.53	0.25	0.04
	Difference between subgroups			-465*	0.01	0.08	0.25	0.58	0.42	0.14	0.02	0.00
LIFT	Not employed in the past month	293	226	43	0.00	0.00	0.03	0.17	0.83	0.52	0.21	0.01
	Employed in the past month	280	1,470	39	0.00	0.00	0.04	0.20	0.80	0.50	0.23	0.02
	Difference between subgroups			4	0.01	0.07	0.22	0.49	0.51	0.23	0.07	0.00
MyGoals	Not employed in the past month	679	211	38	0.00	0.00	0.02	0.16	0.84	0.46	0.13	0.00
	Employed in the past month	568	568	33	0.00	0.00	0.02	0.18	0.82	0.49	0.17	0.00
	Difference between subgroups			5	0.00	0.04	0.19	0.52	0.48	0.17	0.03	0.00

Sources: Evaluation of Employment Coaching baseline and first follow-up surveys and MyGoals Baseline Questionnaire data.

Note: Outcomes are measured over the first 9 months after study enrollment for FaDSS, Goal4 It!, and LIFT. Outcomes are measured over the first 12 months after study enrollment for MyGoals. Estimated impacts and level of statistical significance come from the standard approach to estimating subgroup impacts based on the null hypothesis testing framework.

***/**/* Impact estimates are statistically significant at the .01/.05/.10 levels, respectively, using a two-tailed t-test.

Table B.6. Impacts of all four coaching programs on average monthly self-reported earnings by education level at baseline based on the BHLM approach

					Probability that the true impact or difference in true impacts was...							
Program	Subgroup	Sample size	Control group mean	Estimated impact or difference and level of significance	Less than -$100	Less than -$50	Less than -$25	Less than $0	Greater than $0	Greater than $25	Greater than $50	Greater than $100
(a) Estimating separately for each program												
FaDSS	No college	318	635	54	0.00	0.00	0.06	0.30	0.70	0.33	0.10	0.00
	Some college or higher	179	663	96	0.00	0.01	0.07	0.27	0.73	0.41	0.16	0.01
	Difference between subgroups			-42	0.00	0.05	0.22	0.56	0.44	0.13	0.02	0.00
Goal4 It!	No college	216	478	193*	0.00	0.00	0.03	0.17	0.83	0.55	0.28	0.03
	Some college or higher	223	964	61	0.00	0.00	0.04	0.21	0.79	0.50	0.24	0.03
	Difference between subgroups			132	0.01	0.07	0.20	0.45	0.55	0.27	0.10	0.01
LIFT	No college	372	593	38	0.00	0.00	0.07	0.33	0.67	0.32	0.09	0.00
	Some college or higher	202	1,326	0	0.00	0.02	0.11	0.35	0.65	0.35	0.14	0.01
	Difference between subgroups			38	0.00	0.07	0.23	0.50	0.50	0.22	0.06	0.00
MyGoals	Some college or higher	758	345	6	0.00	0.00	0.06	0.33	0.67	0.26	0.05	0.00
	No college	493	433	72	0.00	0.00	0.04	0.24	0.76	0.40	0.12	0.00
	Difference between subgroups			-66	0.00	0.06	0.26	0.63	0.37	0.10	0.01	0.00

(continued)

Program	Subgroup	Sample size	Control group mean	Estimated impact or difference and level of significance	Probability that the true impact or difference in true impacts was... Less than -$100	Less than -$50	Less than -$25	Less than $0	Greater than $0	Greater than $25	Greater than $50	Greater than $100
(b) Combining information across programs												
FaDSS	No college	318	635	54	0.00	0.00	0.03	0.21	0.79	0.43	0.14	0.00
	Some college or higher	179	663	96	0.00	0.00	0.03	0.18	0.82	0.51	0.21	0.01
	Difference between subgroups			-42	0.00	0.05	0.22	0.57	0.43	0.12	0.02	0.00
Goal4 It!	No college	216	478	*193	0.00	0.00	0.01	0.12	0.88	0.62	0.31	0.03
	Some college or higher	223	964	61	0.00	0.00	0.02	0.14	0.86	0.60	0.30	0.04
	Difference between subgroups			132	0.01	0.08	0.22	0.48	0.52	0.24	0.08	0.00
LIFT	No college	372	593	38	0.00	0.00	0.04	0.22	0.78	0.44	0.15	0.01
	Some college or higher	202	1,326	0	0.00	0.01	0.05	0.22	0.78	0.48	0.21	0.02
	Difference between subgroups			38	0.01	0.09	0.25	0.54	0.46	0.20	0.05	0.00
MyGoals	No college	758	345	6	0.00	0.00	0.02	0.21	0.79	0.39	0.09	0.00
	Some college or higher	493	433	72	0.00	0.00	0.03	0.18	0.82	0.50	0.19	0.01
	Difference between subgroups			-66	0.00	0.05	0.24	0.60	0.40	0.11	0.01	0.00

Sources: Evaluation of Employment Coaching baseline and first follow-up surveys and MyGoals Baseline Questionnaire data.

Note: Outcomes are measured over the first 9 months after study enrollment for FaDSS, Goal4 It!, and LIFT. Outcomes are measured over the first 12 months after study enrollment for MyGoals. Estimated impacts and level of statistical significance come from the standard approach to estimating subgroup impacts based on the null hypothesis testing framework.

***/**/* Impact estimates are statistically significant at the .01/.05/.10 levels, respectively, using a two-tailed t-test.

Table B.7. Impacts of all four coaching programs on average monthly self-reported earnings by number of children at baseline based on the BHLM approach

					Probability that the true impact or difference in true impacts was...							
Program	Subgroup	Sample size	Control group mean	Estimated impact or difference and level of significance	Less than -$100	Less than -$50	Less than -$25	Less than $0	Greater than $0	Greater than $25	Greater than $50	Greater than $100
(a) Estimating separately for each program												
FaDSS	Fewer than two children	176	613	13	0.00	0.01	0.09	0.35	0.65	0.33	0.11	0.01
	Two or more children	321	665	93	0.00	0.00	0.07	0.33	0.67	0.31	0.09	0.00
	Difference between subgroups			-80	0.00	0.02	0.15	0.49	0.51	0.17	0.03	0.00
Goal4 It!	Fewer than two children	194	504	347**	0.00	0.00	0.04	0.20	0.80	0.52	0.24	0.03
	Two or more children	242	941	-131	0.00	0.01	0.07	0.30	0.70	0.38	0.14	0.01
	Difference between subgroups			478**	0.00	0.02	0.11	0.35	0.65	0.31	0.10	0.01
LIFT	Fewer than two children	155	1,207	-7	0.00	0.02	0.11	0.35	0.65	0.35	0.13	0.01
	Two or more children	419	727	30	0.00	0.00	0.06	0.34	0.66	0.30	0.08	0.00
	Difference between subgroups			-37	0.00	0.05	0.19	0.49	0.51	0.21	0.06	0.00
MyGoals	Fewer than two children	637	298	79	0.00	0.00	0.03	0.21	0.79	0.41	0.11	0.00
	Two or more children	583	472	-21	0.00	0.01	0.09	0.38	0.62	0.25	0.05	0.00
	Difference between subgroups			100	0.00	0.01	0.08	0.33	0.67	0.31	0.08	0.00

(continued)

Program	Subgroup	Sample size	Control group mean	Estimated impact or difference and level of significance	Probability that the true impact or difference in true impacts was... Less than -$100	Less than -$50	Less than -$25	Less than $0	Greater than $0	Greater than $25	Greater than $50	Greater than $100
(b) Combining information across programs												
FaDSS	Fewer than two children	176	613	13	0.00	0.00	0.04	0.20	0.80	0.48	0.19	0.01
	Two or more children	321	665	93	0.00	0.00	0.04	0.27	0.73	0.37	0.10	0.00
	Difference between subgroups			-80	0.00	0.01	0.09	0.37	0.63	0.25	0.05	0.00
Goal4 It!	Fewer than two children	194	504	*347**	0.00	0.00	0.02	0.12	0.88	0.61	0.31	0.03
	Two or more children	242	941	-131	0.00	0.00	0.04	0.22	0.78	0.44	0.17	0.01
	Difference between subgroups			478**	0.00	0.02	0.09	0.33	0.67	0.33	0.11	0.01
LIFT	Fewer than two children	155	1,207	-7	0.00	0.00	0.04	0.18	0.82	0.53	0.24	0.02
	Two or more children	419	727	30	0.00	0.00	0.04	0.23	0.77	0.40	0.12	0.00
	Difference between subgroups			-37	0.00	0.02	0.12	0.38	0.62	0.28	0.08	0.01
MyGoals	Fewer than two children	637	298	79	0.00	0.00	0.01	0.13	0.87	0.53	0.18	0.00
	Two or more children	583	472	-21	0.00	0.00	0.05	0.27	0.73	0.36	0.09	0.00
	Difference between subgroups			100	0.00	0.01	0.08	0.33	0.67	0.31	0.07	0.00

Sources: Evaluation of Employment Coaching baseline and first follow-up surveys and MyGoals Baseline Questionnaire data.

Note: Outcomes are measured over the first 9 months after study enrollment for FaDSS, Goal4 It!, and LIFT. Outcomes are measured over the first 12 months after study enrollment for MyGoals. Estimated impacts and level of statistical significance come from the standard approach to estimating subgroup impacts based on the null hypothesis testing framework.

***/**/* Impact estimates are statistically significant at the .01/.05/.10 levels, respectively, using a two-tailed t-test.

Table B.8. Results from simulation of BCF for varying sample sizes and number of variables used to define subgroups (assuming a normal distribution of outcomes)

Number of candidate variables BCF used to define subgroups	Percentage of times BCF identified a likely subgroup difference (at least a 95 percent chance that the impacts differed between two subgroups)
Sample size of 2,000	
5	19
10	14
20	9
Sample size of 5,000	
5	80
10	67
20	55
Sample size of 10,000	
5	98
10	95
20	95
Sample size of 20,000	
5	100
10	100
20	100

Note: The estimates are based on (1) 100 simulation draws for each covariate and sample size combination, (2) an assumed normal distribution of earnings, and (3) a 0.20 effect size difference in impacts across one of the variables used to define the subgroups.


OPRE


Mathematica
Progress Together