MDRC Working Papers on Research Methodology

When Is the Story in the Subgroups?

Strategies for Interpreting and Reporting Intervention Effects on Subgroups

Howard S. Bloom
Charles Michalopoulos

April 2010

This paper was supported by funding from the W.T. Grant Foundation and the Judith Gueron Fund for Methodological Innovation in Social Policy Research at MDRC.

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

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

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

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

Abstract

This paper examines strategies for interpreting and reporting estimates of intervention effects for subgroups of a study sample. Specifically, the paper considers: why and how subgroup findings are important for applied research, the importance of pre-specifying subgroups before analyses are conducted, the importance of using existing theory and prior research to distinguish between subgroups for whom study findings are confirmatory (hypothesis testing), as opposed to exploratory (hypothesis generating), and the conditions under which study findings should be considered confirmatory based on their pre-specification and pattern of statistical significance for the full sample, its subgroups, and their differences. These issues are illustrated by empirical examples from past work by the authors.

Contents

List of Tables

Table

Introduction

In much empirical research, there is interest not only in an overall effect, but also in effects for different groups. For example, Michalopoulos and Schwartz (2000) estimate the effects of a number of welfare-to-work programs on a range of subgroups defined by educational level, employment experience, risk of depression, and so on, with the goal of helping welfare administrators target the right services to their clients. In examining the effects of a transitional jobs program designed to help people who are leaving prison return to work, Bloom, Redcross, Zweig, and Azurdia (2007) found the effects were concentrated among those who had most recently left prison. A recent publication garnered attention in the popular press by providing evidence that antidepressants were effective only for people with severe depression.[1] But how much importance should researchers place on subgroup findings when interpreting and reporting estimates of intervention effects? The goal of this paper is to articulate a strategy for determining how much. We first identify the factors that should determine how subgroup findings are handled. We then summarize several scenarios in which some of the key factors vary and discuss how these factors should help determine which conclusions to draw about subgroup findings.

The audience for this paper includes anyone doing research on the effects of interventions. For policy researchers writing reports for federal and state policymakers, the paper provides some guidelines about which subgroup findings should be highlighted in, for example, an executive summary of a report. For academic researchers, the paper can be used to help decide whether a subgroup finding should be stressed in describing results in a journal article.

We assume that subgroups can be defined in terms of demographic differences (for example, with respect to age, race, gender), geographic differences (for example, with respect to study sites or administrative districts), or temporal differences (for example, with respect to varying follow-up periods) among sample observations. We also present our argument in the context of a random assignment study, although the logic applies equally to nonexperimental analyses.

Based on our reading of the relevant literature, we propose that the following factors should determine the ways in which estimates of intervention effects for subgroups are treated:

1. **Pre-specification of the subgroup.** To discourage researchers from fishing for results, we suggest that subgroups be highlighted only if they were specified before the analysis began, preferably based on theory and/or prior research.

[1] Fournier et al. (2010).

2. **Statistical significance of the subgroup's estimated intervention effect.** We propose that subgroups should not receive much attention if the estimated effect of the intervention for that subgroup is not statistically significant. In that case, the most that can be said is that the study did not provide evidence of an intervention effect for that subgroup.

3. **Statistical significance of subgroup differences in the estimated intervention effect.** We propose that the main question in looking at subgroups should be whether the intervention was significantly more effective for one group than another. If not, we recommend using the overall sample results to describe the effects of the intervention.

4. **Statistical significance of the overall average estimated intervention effect for the study sample.** We propose giving more credence to subgroup differences when the overall effect of the intervention is statistically significant.

5. **Internal contextual factors (that is, the observed pattern of estimated intervention effects across subgroups, outcomes, and/or time points).** Subgroup differences should be treated with greater confidence when the pattern of other effects is consistent with that subgroup difference, but treated with more skepticism when the pattern of other effects is not consistent with the subgroup finding.

6. **External contextual factors (that is, pre-existing theory and/or empirical findings).** Subgroup differences should be treated with greater confidence when external considerations such as theory and prior research are consistent with the subgroup finding, but treated with more skepticism when they are not.

Underlying our approach is an attempt to reduce the possibility that chance findings will be emphasized: hence our focus on statistical significance — to ensure that findings are unlikely to be due to chance — and internal and external contextual factors — to raise skepticism about results that are statistically significant but not in accordance with other considerations.

We propose two categories of estimated intervention effects: (1) exploratory findings and (2) confirmatory findings. Exploratory findings provide a basis for developing hypotheses that can be tested by future research. Such findings should be considered suggestive only and do not provide a basis for testing hypotheses. In contrast, confirmatory findings provide a basis for testing hypotheses. If consistent with theory, statistically significant, large in magnitude, and not sensitive to variations in estimation methods and sample definition, confirmatory findings should be considered strong evidence of an intervention's effect or lack thereof. We therefore

conclude that confirmatory findings have a legitimate place in executive summaries and key chapters of reports for policymakers and deserve prominent discussion in empirical journal articles. In contrast, exploratory findings should be less prominently displayed and discussed. Hence they should be considered in chapters of an evaluation report or sections of a journal article that are more speculative than definitive.

In the sections that follow, we consider how each of the preceding factors affects whether a subgroup finding should be considered exploratory or confirmatory. In so doing, we identify those points about which we expect general agreement in the research community and those points where disagreement is more likely. We also note several points that bear directly on how to handle subgroup findings but that have not yet been discussed systematically. In closing, we briefly raise the problem of multiple-hypothesis testing, which lies at the heart of controversies over how to handle subgroup findings.

Pre-Specification

In the existing literature — especially that on medical research — pre-specifying a subgroup finding is regarded as an indispensable condition for producing serious evidence.[2] Pre-specification might be based on existing theory about how the defining feature of a subgroup (such as the severity of an existing condition) interacts with the intervention to be tested, or on established empirical evidence about how the subgroup's reaction to a similar intervention differs from that of other subgroups. Both of these information sources can provide a legitimate and plausible rationale for expecting an intervention to affect a given group differently from other groups. The stronger the pre-existing information is, the stronger the subsequent finding will be, if the hypothesized subgroup result is observed. This process of accumulating theoretical and empirical evidence lies at the heart of the modern scientific method.

Another source of interest in subgroup findings, and hence, their pre-specification, is policy relevance or political salience. This is a particularly important impetus for examining findings for many of the subgroups that play key roles in reports intended for policymakers. It is less clear, however, whether this rationale should have the same scientific status as does pre-existing theory or empirical findings.

Our first recommendation is that subgroup findings should not be considered as confirmatory if they were not specified in advance of the analysis for the report or article in which they are presented. Pre-specification should be done as early as possible during the design or implementation of a study. There are particular advantages to specifying the subgroups while

[2]For example, see Rothwell (2005).

the study is being designed in order to ensure the study has enough statistical power to detect the relevant subgroup differences. At a minimum, however, the subgroups should be specified before any analysis is done.

Statistical Significance

Most of our discussion about how to handle subgroup findings focuses on alternative configurations of statistical significance for those findings. In this regard, it is first necessary to determine whether a specific subgroup finding is itself statistically significant. If not, then all that can be said is that the study does not provide evidence of an intervention effect for the particular subgroup. However, it is important to note that neither does such a finding provide evidence of the lack of an intervention effect for the subgroup.

If the estimated intervention effect for the subgroup is statistically significant, then its confirmatory versus exploratory status must be judged in the context of the significance of findings for other subgroups and/or the full study sample. In particular, when an estimated effect for a subgroup is significantly different from zero, it is important to consider whether the estimated intervention effect for the subgroup differs statistically significantly from the corresponding result for the rest of the study sample.

When Impact Estimates Significantly Differ by Group

If the difference between a subgroup and the rest of a sample is statistically significant (and the subgroup finding was pre-specified), then it should be considered confirmatory and can be highlighted. However, meeting this condition is challenging, because of the typically limited power of statistical tests of group differences in intervention effects. For example, with two subgroups of equal size, the minimum detectable difference between their estimated intervention effects is twice the magnitude of the minimum detectable intervention effect for their combined sample. Hence, it is often the case that seemingly large subgroup differences in estimated intervention effects are not statistically significant. This is one reason to specify the key subgroups before the study begins: Doing so helps to ensure that the study has enough statistical power to detect important subgroup differences.

To provide an example of a situation in which a study produces significant subgroup differences, we use results from a study of the Working toward Wellness program. This intervention is being studied at the Rhode Island site of the Enhanced Services for the Hard-to-Employ evaluation, which is being funded primarily by the Administration for Children and Families and the Office of the Assistant Secretary for Planning and Evaluation in the U.S.

Department of Health and Human Services.[3] In Rhode Island, parents (mostly mothers) receiving Medicaid were recruited into the study if they appeared to be depressed based on a standard set of interview questions. Half were randomly placed into the program group, which received outreach from Master's level clinicians who encouraged participants to seek treatment for their depression and who monitored their treatment. The other half were randomly placed in a control group, which could use any services available to other Medicaid recipients in Rhode Island. Six months after random assignment, individuals were interviewed and administered a set of questions to assess the severity of their depression. In addition, medical claims data were available for all sample members from the managed care organization providing Medicaid services in Rhode Island.

Before the impact analysis was conducted, but after the study was designed and the sample enrolled, the study team chose two sets of subgroups to analyze. One set of subgroups was based on whether individuals were Hispanic or non-Hispanic and was chosen by the study team because a prior study had found that outreach to engage individuals in treatment for depression had larger effects for Hispanics than for others. The second set of subgroups was defined based on individuals' depression severity at baseline, and was chosen based on prior research and advice from a psychiatrist advising the study team.

Table 1 shows results for the full sample and the two subgroups for two outcomes: (1) the proportion of sample members who filled a prescription for an antidepressant during the six months following random assignment and (2) the average depression severity score from the six-month follow-up survey. Although the estimated effect was not statistically significant for the full sample, the impacts were significantly larger for Hispanic sample members than for others (as indicated by the dagger symbol in the table). While the program had essentially no effect on the proportion of non-Hispanic sample members taking antidepressants or on their average depression score, it increased the proportion of Hispanic sample members taking antidepressants by 14.3 percentage points and reduced their average depression severity by 2.3 points on a 30-point scale.

When Impact Estimates Do Not Significantly Differ by Group

When estimated effects are not statistically significantly different between a subgroup and the rest of the sample, the next step is to look at the statistical significance of estimated intervention effects for the full study sample and other subgroups. Here we consider four cases, depending on whether the impact estimate for the full sample is statistically significant and whether impact estimates for other subgroups are statistically significant.

[3]Kim, LeBlanc, and Michalopolous (2009).

Case 1: All impact estimates are statistically significant. The simplest situation to interpret is when the estimated effect for the full sample is significantly different from zero and all subgroup estimates are statistically significant and in the same direction. In this case, the finding that the intervention affected all subgroups could be considered confirmatory. For example, if estimated intervention effects were statistically significantly positive for men and women separately and together, the finding of effectiveness for men (or women) would be confirmatory if it was pre-specified. However, it would be inappropriate to emphasize the results for any particular subgroups since the evidence indicates it is effective for all of them, and the lack of a statistically significance difference between the subgroups (no daggers in the table) suggests that any observed differences are likely due to chance.

Table 2 shows an example of this case from the Rhode Island study. In this case, medical claims data were used to calculate the number of visits each person made to a mental health professional, such as a psychiatrist, psychologist, or counselor, in the six months following random assignment. Because getting people into mental health treatment was the direct goal of the intervention, it is not surprising that the program increased the average number of mental health visits for the full sample and for both the Hispanic and non-Hispanic subgroups. The study team therefore felt comfortable concluding that the program was successful in this respect for both subgroups.

Case 2: No impact estimates are statistically significant. A related scenario is when estimated intervention effects are not statistically significant for any related subgroups or for the overall study sample. In this case, the most that can be said about a subgroup of interest is that the study did not find convincing evidence of an intervention effect for it.

Table 3 shows an example of this case from Rhode Island. The outcome is the proportion of sample members who received antidepressants over the first 18 months of follow-up, which included six months after the intervention had ended. In this case, the medical claims data found small and statistically insignificant results for the full sample and each subgroup.[4] Combining this finding with the six-month finding on antidepressants shown in Table 1, the study team concluded that the intervention produced larger effects on use of antidepressants for Hispanic sample members while the intervention was still ongoing, but that that difference disappeared after individuals stopped receiving the intervention. This was confirmed by the pattern of impacts for several other types of health care use, which showed robust effects during the year of the intervention and then the disappearance of those effects after the intervention's end.

[4]Kim et al. (forthcoming).

Case 3: Impact estimates statistically significant for only one subgroup. A third related scenario for impact estimates that do not differ across subgroups is when estimated intervention effects are statistically significant for a subgroup of interest but not statistically significant for the rest of the study sample or for the full study sample. This might have occurred in the Rhode Island study, for example, if the estimated intervention effect were statistically significantly positive for the Hispanic subsample but not statistically significant for non-Hispanic sample members or for the full sample. In this case, we recommend that the results be considered exploratory and not highlighted. The rationale for this decision is that in the absence of convincing information to the contrary (such as a statistically significant difference among subgroup findings) the best information about findings for a subgroup is the corresponding result for the full study sample.

Since there were no instances of this case in the Rhode Island study, Table 4 shows an example from the evaluation of the Center for Employment Opportunity's (CEO) transitional jobs program for men leaving prison.[5] The study included men who had been recently released from prison and who lived in New York City. The randomized program group was placed into subsidized jobs for up to six months with the goal of helping them find unsubsidized employment before the six-month period was out. The randomized control group received resources to help them look for work, but did not have access to the subsidized jobs.

The literature on helping men avoid returning to prison suggests that it is important to intervene soon after a person has been released from prison, if not before. The study team therefore expected the study sample to include men who had been released quite recently, within a month or two. However, once they examined the data for men who had been referred to CEO for services and randomly assigned into the study, they discovered that a sizable group had been released from prison three or more months prior to entering the study. To test the hypothesis that programs such as this would be more effective for recently released prisoners, they therefore divided the sample roughly in half into a group that entered the study within three months of having been released from prison and a group that had been out of prison longer than three months before they entered the study.

Table 4 shows estimated effects of the program on the full sample and the two subgroups on one measure of recidivism, namely the proportion of sample members who were arrested, convicted, or re-incarcerated in the year after random assignment. The results all suggest that CEO modestly reduced recidivism, but the estimated effect for the full sample was not statistically significant, nor was the estimated effect for the group that had not been recently

[5]The results shown in Table 4 are unpublished, but published findings are available from Bloom, Redcross, Zweig, and Azurdia (2007) and Redcross et al. (2009).

released. Although the estimated effect for the recently released subgroup was statistically significant at the 10 percent level, the difference in estimated impacts between the two groups was not statistically significant, suggesting that the effect might have been the same for the two groups. The study team therefore concluded that there was not convincing evidence that CEO had a greater effect on this measure of recidivism for the recently released subgroup and that this result should not be highlighted. It is worth noting, however, that other estimated effects were significantly larger for the recently released subgroup, suggesting that the study team's hypothesis was correct. That hypothesis is being further tested in a six-site study of transitional jobs for reentering prisoners in several Midwestern states.

Case 4: Impact estimates statistically significant for one subgroup and the full sample. The fourth scenario may be the most controversial. It occurs when the finding for the subgroup of interest is statistically significant, the corresponding finding for the full study sample is in the same direction and is statistically significant, but the corresponding finding for the rest of the study sample is not statistically significant.

Table 5 illustrates this case using six-month results from the Rhode Island study. Here, the outcome is the proportion of sample members who received mental health services during the six months following random assignment. Recall from Table 2 that there were statistically significant estimated effects on the number of mental health visits for the full sample and each subgroup, but that differences between the subgroups were not statistically significant. In that case, we concluded that the intervention appeared to increase the number of mental health visits for both subgroups, but not more for one than the other.

According to Table 5, the program increased the proportion of sample members who received any mental health services by 10.5 percentage points, which was significant at the 1 percent level. It also increased the proportion of Hispanic sample members who received any mental health services by 17.6 percentage points, and that estimate was significant at the 5 percent level. While the estimated effect on non-Hispanic sample members was 5.4 percentage points, that estimate was not statistically significant. Moreover, the difference in estimated effects was not significantly larger for the Hispanic subgroup than for the non-Hispanic subgroup.

The question in this case is whether to conclude that the program benefited Hispanic sample members while saying nothing about non-Hispanic sample members or to conclude that the program had widespread effects. The former conclusion is based on the fact that the estimate for only one of the subgroups is statistically significant. The latter conclusion is based on the finding that the estimated effects for the two groups did not significantly differ and that findings for the full sample were statistically significant. Here are the two possible positions stated more generally.

Position A: The finding for the subgroup of interest (Hispanic sample members in the Rhode Island example) is confirmatory (assuming that the subgroup distinction was pre-specified) because it is statistically significant in its own right and is consistent with the best information for that subgroup, absent direct information for it (the corresponding result for the full study sample). This finding does not mean that the study found no intervention effect for the rest of the study sample (non-Hispanic sample members in the Rhode Island example). The most that can be said about this residual subgroup is that the study did not find direct evidence of an intervention effect for it, although there was indirect evidence from results for the full study sample.

Position B: The finding for the subgroup of interest (Hispanic sample members in the example) is exploratory because there was no statistically significant difference between findings for the subgroup and the rest of the study sample, and to advertise the significant finding for the subgroup of interest makes it look (by comparison) like the study found no intervention effect for the rest of the sample. In other words, this encourages invidious comparisons among the subgroup findings.

To some extent, the two positions are based on different rationales for examining subgroups and differing views about the importance of estimated effects for the full sample. Proponents of Position A are probably most interested in examining results by subgroup to be able to make statements about a subgroup, regardless of how it compares to other subgroups. Proponents of Position B probably focus more on how different subgroups compare to one another. If differences among them are not statistically significant, they choose to highlight the overall study effects rather than relying on results that could easily be due to chance.

Contextual Considerations

Two additional factors should affect how subgroup findings are considered: internal and external contextual considerations. We include these points to acknowledge the importance of interpreting all scientific findings in their relevant contexts.

By internal contextual considerations we mean features of findings that are internal to a given study. For example, it is often argued that a *pattern* of findings can provide important evidence about intervention effects even if the separate findings involved are not statistically significant and thus cannot stand on their own. Common examples of such patterns include consistently positive estimates of intervention effects across related outcome measures and/or over time during a follow-up period. In the Rhode Island example, there were a number of significant differences in impact estimates between Hispanic and non-Hispanic sample members during the six months following random assignment, and that pattern gave the team more confidence that there was a true difference. By 18 months after random assignment, there were

few differences between the two groups. Moreover, the reduction in impacts coincided with the program's end after one year. This gave the team confidence that, for Hispanic sample members, the program had only temporary effects that disappeared after the program ended.

By external contextual considerations we mean features of findings that are external to a given study. For example, results that are consistent with prior research should be treated with more confidence than results that contradict prior research. Likewise, results that are consistent with a well-recognized or well-conceived theory should receive more prominent attention. By adding either or both of these components, one can interpret findings from a given study in a broad context.

Multiple Hypothesis Testing

In closing, we feel obliged to raise the specter of distortions to statistical inferences that occur when multiple tests are conducted. This issue of multiple testing, as it's called, has been largely ignored in past intervention studies but has risen to the fore in recent years. There appear to be four main approaches to minimizing the risk of incorrectly concluding that specific estimates of intervention effects are real when they appear to be statistically significant in the context of many tests. One approach, which we fully endorse, is to explicitly distinguish between exploratory and confirmatory findings. A second approach, which we also endorse, is to minimize the number of confirmatory hypothesis tests conducted for a given study. These decisions should be made well before any analyses are conducted for a given study and, if possible, during the development of a project's proposal or design paper. We believe that the benefits of carefully making tradeoffs among competing research questions early in the development of a study can be huge.

A third approach to protecting against incorrect statistical inferences is to add an omnibus test that considers all outcome measures and subgroups at the same time. A popular version of this approach is to test the statistical significance of the estimated impact on a composite measure of individual outcomes for all subgroups together (the full study sample). If the estimated impact of the intervention on this composite outcome for the full study sample is statistically significant then the composite test helps to add confidence to separate tests for individual outcome measures and for subgroups of the sample. If the overall composite test does not indicate a significant composite intervention effect, it calls into question whether significant findings for specific outcome measures and/or subgroups are real. Unfortunately, this approach has a number of limitations that fall outside of the purview of this paper.

A fourth approach to guarding against incorrect statistical inferences is to make adjustments (such those named after Bonferonni or Benjamini and Hochberg) to the level of statistical significance (p-value) for each individual hypothesis test. Unfortunately, this approach typically

overcompensates for multiple testing and thus wastes already limited statistical power for estimates of intervention effects. Because of this problem, we have not used this approach in our own research and we are reluctant to recommend it to others.

Conclusion

The goal of this paper is to propose a set of criteria to help researchers decide whether, when, and upon what basis to highlight subgroup findings. The overarching goal of our proposed criteria is to attempt to reduce the likelihood of highlighting chance findings and increase the likelihood of highlighting findings of true policy relevance. For that reason, we give special prominence to subgroup differences that reach the threshold of statistical significance, because no other measure suggests that findings are unlikely to be due to chance. At the same time, to avoid the possibility of looking at many subgroups until an interesting finding emerges, we recommend that subgroups be specified before the analysis begins. Likewise, contextual factors are important because subgroup findings that jibe with internal and external contexts are more likely to be true.

Table 1

Significant Differences Between Subgroups

Hispanic and Non-Hispanic Sample Members in the Working toward Wellness Study Six Months After Random Assignment

Outcome	Program Group	Control Group	Difference (Impact)	P-Value
Filled a prescription for an antidepressant (%)				
Full sample	38.5	34.5	4.0	0.299
Hispanic subgroup	43.7	29.3	14.3 *	0.055 †
Non-Hispanic subgroup	36.2	36.5	-0.3	0.956 †
Depression severity (30-point scale)				
Full sample	12.5	12.8	-0.4	0.51
Hispanic subgroup	12.6	14.9	-2.3 **	0.05 ††
Non-Hispanic subgroup	12.4	12.0	0.4	0.53 ††

SOURCE: Information on antidepressants was calculated from United Behavioral Health medical claims data. Information on depression severity was based on a six-month follow-up survey.

NOTES: Statistical significance levels for the full sample and individual subgroups are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent. Statistically significant differences between the Hispanic and non-Hispanic subgoup are indicated as ††† = 1 percent; †† = 5 percent; and † = 10 percent.

Table 2

Significant Impacts for Both Subgroups and Full Sample

Hispanic and Non-Hispanic Sample Members in the Working toward Wellness Study Six Months After Random Assignment

Outcome	Program Group	Control Group	Difference (Impact)	P-Value
Number of visits to a mental health professional				
Full sample	2.3	1.1	1.2 **	0.017
Hispanic subgroup	2.7	0.9	1.8 **	0.012
Non-Hispanic subgroup	1.7	1.0	0.7 *	0.092

SOURCE: Information on antidepressants was calculated from United Behavioral Health medical claims data.

NOTES: Statistical significance levels for the full sample and individual subgroups are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent. Differences in impacts between the Hispanic and non-Hispanic subgoups were not statistically significant.

Table 3

Insignificant Impacts for Both Subgroups and Full Sample

Hispanic and Non-Hispanic Sample Members in the Working toward Wellness Study Eighteen Months After Random Assignment

Outcome	Program Group	Control Group	Difference (Impact)	P-Value
Filled a prescription for an antidepressant (%)				
Full sample	52.8	49.5	3.3	0.418
Hispanic subgroup	53.8	47.1	6.7	0.383
Non-Hispanic subgroup	52.1	50.6	1.5	0.770

SOURCE: Information on antidepressants was calculated from United Behavioral Health medical claims data.

NOTES: Statistical significance levels for the full sample and individual subgroups are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent. Differences in impacts between the Hispanic and non-Hispanic subgoups were not statistically significant.

Table 4

Significant Impacts for Only One Subgroup

Evaluation of the Center for Employment Opportunity's Transitional Jobs Program

Outcome	Program Group	Control Group	Difference (Impact)
Arrested, convicted, or incarcerated (%)			
Full sample	47.1	51.8	-4.7
Released from prison less than 3 months before entering study	50.3	58.8	-8.5 *
Released from prison more than 3 months before entering study	45.9	46.2	-0.3

SOURCE: Information from New York State criminal justice records.

NOTES: Statistical significance levels for the full sample and individual subgroups are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent. Differences in estimated impacts between the two subgroups were not statistically significant.

Table 5

Significant Impacts for One Subgroup and the Full Sample

Hispanic and Non-Hispanic Sample Members in the Working toward Wellness Study Eighteen Months After Random Assignment

Outcome	Program Group	Control Group	Difference (Impact)	P-Value
Received mental health services (%)				
Full sample	32.2	21.7	10.5 ***	0.007
Hispanic subgroup	39.2	21.6	17.6 **	0.019
Non-Hispanic subgroup	27.7	22.4	5.4	0.268

SOURCE: Information was calculated from United Behavioral Health medical claims data.

NOTES: Statistical significance levels for the full sample and individual subgroups are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent. Differences in estimated impacts between the Hispanic and non-Hispanic subgoups were not statistically significant.

References

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

Fournier, Jay C., Robert J. DeRubeis, Steven D. Hollon, Sona Dimidjian, Jay D. Amsterdam, Richard C. Shelton, and Jan Fawcett. 2010. "Antidepressant Drug Effects and Depression Severity: A Patient-Level Meta-analysis." *Journal of the American Medical Association* 303(1): 47-53.

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

Kim, Sue, Allen Leblanc, Greg Simon, Johanna Walter, and Charles Michalopoulos. Forthcoming. *Working toward Wellness: Eighteen-Month Results from a Telephonic Care Management Program for Medicaid Recipients with Depression*. New York: MDRC.

Michalopoulos, Charles, and Christine Schwartz. 2000. *What Works Best for Whom: Impacts of 20 Welfare-to-Work Programs by Subgroup*. Washington, DC: U.S. Department of Health and Human Services, Office of the Assistant Secretary for Planning and Evaluation and Administration for Children and Families, and U.S. Department of Education.

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

Rothwell, Peter M. 2005. "Subgroup Analysis in Randomised Control Trials: Importance, Indications and Interpretation." *The Lancet* 365: 176-86.

About MDRC

MDRC is a nonprofit, nonpartisan social policy research organization dedicated to learning what works to improve the well-being of low-income people. Through its research and the active communication of its findings, MDRC seeks to enhance the effectiveness of social and education policies and programs.

Founded in 1974 and located in New York City and Oakland, California, MDRC is best known for mounting rigorous, large-scale, real-world tests of new and existing policies and programs. Its projects are a mix of demonstrations (field tests of promising new program approaches) and evaluations of ongoing government and community initiatives. MDRC's staff bring an unusual combination of research and organizational experience to their work, providing expertise on the latest in qualitative and quantitative methods and on program design, development, implementation, and management. MDRC seeks to learn not just whether a program is effective but also how and why the program's effects occur. In addition, it tries to place each project's findings in the broader context of related research — in order to build knowledge about what works across the social and education policy fields. MDRC's findings, lessons, and best practices are proactively shared with a broad audience in the policy and practitioner community as well as with the general public and the media.

Over the years, MDRC has brought its unique approach to an ever-growing range of policy areas and target populations. Once known primarily for evaluations of state welfare-to-work programs, today MDRC is also studying public school reforms, employment programs for ex-offenders and people with disabilities, and programs to help low-income students succeed in college. MDRC's projects are organized into five areas:

- Promoting Family Well-Being and Child Development
- Improving Public Education
- Promoting Successful Transitions to Adulthood
- Supporting Low-Wage Workers and Communities
- Overcoming Barriers to Employment

Working in almost every state, all of the nation's largest cities, and Canada and the United Kingdom, MDRC conducts its projects in partnership with national, state, and local governments, public school systems, community organizations, and numerous private philanthropies.