# When Is the Story in the Subgroups?

## Strategies for Interpreting and Reporting Intervention Effects for Subgroups

**Howard S. Bloom**
**Charles Michalopoulos**

**Revised November 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.

Authors' Note: Correspondence concerning this article should be addressed to Howard Bloom, MDRC, 16 East 34th Street, New York, NY 10016. E-mail: howard.bloom@mdrc.org

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

# Abstract

This revised working paper examines strategies for interpreting and reporting estimates of intervention effects for subgroups of a study sample. The paper considers why and how subgroup findings are important for applied research, alternative ways to define subgroups, different research questions that motivate subgroup analyses, and the importance of prespecifying subgroups before analyses are conducted. It also considers 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. Each issue is illustrated by selected empirical examples.

# Table of Contents

# List of Tables

**Table**

## Introduction

In much applied research, there is interest not only in the overall average effect of an intervention, but also in its effects for different subgroups of sample members. For example, Michalopoulos and Schwartz (2000) estimate the effects of welfare-to-work programs on economic outcomes for a broad range of subgroups defined by characteristics, such as education level, prior employment experience, and risk of depression. They conducted their analysis to help welfare administrators better target services to clients. In a second example, Bloom, Redcross, Zweig, and Azurdia (2007) examine the effects of a transitional jobs program for ex-offenders. They found that program effects were concentrated mainly among sample members who had been released from prison most recently. In a third example, Fournier et al. (2010) garnered widespread attention from their empirical evidence that antidepressants are effective only for people with severe depression. In a fourth example, Bloom, Levy Thompson, and Unterman (2010) found that new small high schools of choice in New York City increased progress toward graduation for a wide range of student subgroups defined in terms of socio-economic characteristics and prior educational attainment. But how much importance should researchers place on subgroup findings such as these? Specifically, what subgroup findings should they emphasize and under what conditions should they do so?

The goal of this paper is to outline a strategy for making these decisions. The paper first introduces some of the many different ways that subgroups can be defined, describes the different types of research questions that can motivate subgroup analyses, and identifies key factors that should be used to determine how subgroup findings are reported and interpreted. The paper then describes several scenarios that illustrate how these factors vary in practice and how this variation can influence decisions about reporting subgroup findings.

The audience for the paper includes anyone who is conducting research on the effects of interventions. For policy researchers writing reports for government officials, agency staff members, and the general public, the paper attempts to provide guidance about the types of subgroup findings to highlight in, for example, an executive summary. For academic researchers writing mainly for other researchers in their field, the paper attempts to provide guidance about the types of subgroup findings to highlight in the abstract and conclusion of a journal article.

## Characteristics That Define Subgroups

Subgroups can be defined many different ways. For example, they can be defined in terms of risk factors, such as past smoking, drinking, drug abuse, current health or mental-health status, or the severity of a given disease or problem that an intervention is intended to treat. This

is often done because researchers have reason to believe that responses to the intervention will depend on the severity of the condition it is being used to treat. Subgroups also can be defined in terms of demographic characteristics, such as age, race, or gender, which are believed to be related to the need for a particular intervention or the likelihood of a beneficial response to it. In addition, subgroups can be defined in terms of geographic location or site, such as sample members' municipality, county, or state of residence, or the administrative entity (hospital, welfare office, or school) that is responsible for serving them. Furthermore, subgroups can be defined in terms of the time period during which they were selected for an intervention (their cohort). These last two bases for defining subgroups are often used to reflect differences in the implementation or context of an intervention that are expected to influence its effectiveness.

In addition to defining subgroups in terms of a single characteristic, it is also possible to define them in terms of *combinations* of characteristics. One might, for example, define subgroups for a study of a treatment for depression jointly in terms of sample members' age and current level of depression. Such combinations of observed characteristics are often used to define subgroups in terms of latent or unobserved characteristics. For example, although the true risk of a negative outcome might not be directly observable, it might be possible to measure its predicted risk through a combination of characteristics that are correlated with the outcome.

This paper focuses on single characteristics or combinations of characteristics used to define subgroups that are *exogenous* to the intervention being studied, which means that they are not affected by the intervention or correlated with its receipt. In randomized trials, this is the case for all baseline characteristics, because they are determined before sample members are randomized to treatment or control status. A different type of characteristic that is sometimes used to define subgroups is one that is *endogenous* to the intervention being studied. This means that it is affected by the intervention or correlated with its receipt. For example, researchers sometimes attempt to determine how effects of an intervention vary with differences in the extent or intensity of its receipt (its dosage). Valid causal inferences of this type are much more difficult to make (absent exogenous factors, such as random assignment, that cause the dosage to vary) than are those for differences in intervention effects for subgroups defined by exogenous characteristics. Although a discussion of this added difficulty is beyond the scope of this paper, the issues raised and conclusions drawn by the paper apply with equal force to subgroups defined by endogenous or exogenous characteristics.

## Research Questions That Motivate Subgroup Analyses

In addition to being defined in different ways, subgroups can become the focus of analysis for reasons that reflect different research questions. For example, a researcher might conduct subgroup analyses to address the question:

**How widespread are the effects of an intervention?** Specifically, to what extent are its effects dispersed across many different types of sample members, as opposed to being concentrated within a homogeneous subgroup? To address this question typically requires estimating intervention effects for a wide range of subgroups, comparing the magnitudes of these estimates, and assessing the statistical significance of their differences. To the extent that positive findings are broadly distributed across subgroups (with few large and statistically significant differences), the intervention's effectiveness is robust. To the extent that these effects are highly concentrated among certain subgroups, the intervention's effectiveness is more limited.

**Is the intervention effective for a specific subgroup?** In many fields of study, a specific subgroup is judged to be of special interest because it has a particular need for intervention or because past research has found it to be especially difficult/easy to serve, or both. In these situations it is important to know how well an intervention works for the particular subgroup, almost regardless of its effectiveness for other sample members. Consequently, greater attention is placed on the magnitude of the estimated effect for the subgroup and less attention is placed on its difference from that for other sample members or on findings for the full study sample. For studies like these, special attention is required to ensure adequate statistical power for estimating the subgroup effect of interest.

**Is the intervention effective for any subgroup?** This situation arises when findings for a full study sample suggest that an intervention is not effective on average. Before abandoning the intervention based on this information, it is important to assess whether the full-sample average masks important positive results for a key subgroup. This can occur, for example, when a medical procedure is effective for patients with an early stage of a disease (because they are still in good health) but not effective for patients with an advanced stage of the disease (who are greatly weakened by this point). It is especially important in situations like these to have the discipline of advance planning and structured hypothesis testing in order to reduce the likelihood of subgroup results that reflect merely errors in random sampling, measurement, or estimation.

## Confirmatory and Exploratory Subgroup Findings

Regardless of how subgroups are defined or why they are a focus of analysis, we propose the following framework for making decisions about how to report and interpret their

findings. At the core of this framework is the well known, but often ignored, distinction between *exploratory* and *confirmatory* empirical findings

Exploratory findings provide a basis for generating hypotheses that can be tested rigorously by future research. This essential step in the scientific method provides a basis for developing new theories and extending existing ones. However, exploratory findings should be considered suggestive until they have been replicated by future research. Hence, they should not be used as a basis for testing current hypotheses.

In contrast, confirmatory findings provide an appropriate basis for testing current hypotheses. If they are: (1) produced by a research design that supports valid causal inferences, (2) consistent with existing theory and prior empirical research, (3) statistically significant, (4) large enough in magnitude to be important, and (5) robust to variations in estimation methods and sample definitions, confirmatory findings can provide strong evidence that an intervention is effective. Likewise, if confirmatory findings meet all of the preceding conditions (except perhaps for statistical significance) and have a *narrow confidence interval* around zero effect, they can provide strong evidence that an intervention is *not effective*.

As argued below, we recommend that it is appropriate to display confirmatory subgroup findings prominently in a report or article —- for example, in its abstract, executive summary, and/or concluding section. In contrast, we recommend that although exploratory subgroup analyses should always be reported if they were conducted (to promote full information about the research that was carried out), their findings should not be highlighted *in most cases*. Although it is beyond the scope of the present discussion to explore possible exceptions to this rule, we believe that at a minimum, the burden of proof for emphasizing exploratory subgroup findings should be quite high.[1]

Furthermore, when exploratory subgroup findings are reported, they should include caveats that clearly indicate their current suggestive status. This is not to imply that exploratory findings can never be important. Indeed, a key benefit of systematic scientific inquiry is that it can and does produce unanticipated discoveries. However, the logical process of *exploration*, by which scientific discoveries are made, does not provide the same strength of evidence as does the logical process of *confirmation*, by which scientific hypotheses are tested. And this difference in *strength of evidence* should be made clear to one's readers.

[1] As with any attempt to collapse a continuous construct based on multiple considerations (like strength of scientific evidence) into a dichotomy (confirmatory versus exploratory findings or strong versus weak evidence), it is not possible to distinguish between the resulting categories in a way that fits all possible situations. Hence, in practice, the operational distinction must remain somewhat vague, and its application to specific cases will require professional judgment.

## Factors That Distinguish Between Confirmatory and Exploratory Subgroup Findings

That said, we recommend that a combination of the following factors and the conditions they represent be used to distinguish between confirmatory and exploratory subgroup findings. Meeting any one condition should be a necessary but not sufficient condition for a subgroup finding to be confirmatory. The main purpose for imposing these conditions is to reduce the likelihood of overinterpreting subgroup findings that represent random error.[2]

1. **Prespecification of the subgroups studied.** A basic rule of applied scientific research (that is often honored in the breach) is that only findings for subgroups specified in advance of one's analysis (preferably based on prior theory or empirical research) be eligible for confirmatory status.

2. **Statistical significance of a subgroup's estimated effect.** Only if a subgroup finding is statistically significant should a researcher present it as strong evidence that a treatment had an effect for that subgroup. If the subgroup finding is not statistically significant, the most that can be said on the basis of this fact alone is that the study provides no direct evidence of an effect for the subgroup. This finding does not, however, necessarily indicate that the treatment has no effect or a negligible effect for the subgroup.

3. **Statistical significance of subgroup differences in estimated effects.** Other things (discussed later) being equal, findings for a specific subgroup should not be highlighted unless they differ statistically significantly from those for other sample members. If subgroup differences are not statistically significant, findings for the full study sample usually should be emphasized instead of those for the subgroup.

4. **Statistical significance of the overall average estimated effect for the full-study sample.** Other things being equal, more credence should be given to positive subgroup findings when the estimated full-sample effect is positive and statistically significant than when this is not the case.

5. **Internal contextual factors** (such as the observed pattern of estimated effects across subgroups, outcomes, and/or time points). Subgroup differences should be treated with greater confidence when the pattern of other estimated effects is consis-

[2]Our recommendations are designed to minimize the risk of Type I error in statistical hypothesis testing about intervention effects for subgroups. We recognize that other things being equal, reducing the risk of Type I error (wrongly emphasizing a subgroup finding that is not real or important) increases the risk of Type II error (wrongly not emphasizing a subgroup finding that is real and important). We also acknowledge that there is no general consensus about how to balance the trade-off between these two types of errors.

tent with that subgroup difference and treated with more skepticism when the pattern of other estimated effects is not consistent with the subgroup finding.

6. **External contextual factors** (such as preexisting theory and empirical findings). Subgroup differences should be treated with greater confidence when external considerations, such as preexisting theory and empirical findings, are consistent with the subgroup finding, and treated with more skepticism when they are not.

We now consider how each of the preceding factors can affect whether a subgroup finding should be considered exploratory or confirmatory. In so doing, we identify those points about which we expect general agreement among researchers and those points where we expect disagreement.

### Prespecification

In the existing literature — especially that on medical research — prespecification of a subgroup is regarded as an indispensable condition for findings for that subgroup to provide convincing evidence (see Rothwell, 2005). This prespecification might be based on existing theory about how the defining feature of a subgroup (such as the severity of a preexisting condition) interacts with the intervention being tested, or based on past empirical evidence about how the subgroup's reaction to a similar intervention differs from that of other population members. Both of these information sources can provide a legitimate and plausible rationale for expecting an intervention to affect members of a subgroup differently from others. The stronger this preexisting information is, the stronger the subsequent combined evidence will be if the hypothesized subgroup result is observed. This process of *accumulating* theoretical and empirical evidence by building directly and explicitly on past research is at the core of the scientific method.

Another source of interest in subgroup findings and, hence, a basis for their prespecification 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 preexisting theory or empirical findings.

In light of the importance of prespecification for scientific research, our first recommendation is that subgroup findings should be considered confirmatory only if they were specified in advance of the analysis for the report or article in which they are presented. Such prespecification should be done as early as possible during the design or implementation of a study. There are particular advantages to specifying subgroups while the study is being designed; for example, prespecification helps ensure that appropriate data are collected to identify subgroup members and that the study's sample has enough statistical power to detect relevant

subgroup differences in intervention effects. At a minimum, however, confirmatory subgroups should be specified *before any intervention effects are estimated* for a given study.

### Statistical Significance

If a subgroup is prespecified, one should consider the pattern of statistical significance of findings for the subgroup, for the rest of the sample, and for the full sample in order to determine whether a given study provides confirmatory evidence about the intervention's effectiveness for the subgroup.

The first step in this process is to determine whether the subgroup finding is itself statistically significant. If not, as noted above, the study does not provide direct evidence of an intervention effect for the subgroup. (As noted earlier, this null finding does not demonstrate that the intervention has no effect or a negligible effect for the subgroup. Only if the corresponding estimate has a narrow confidence interval around zero is this negative conclusion warranted.)

If the estimated intervention effect for the prespecified subgroup is statistically significant, its confirmatory status should be judged in the context of the statistical significance of findings for other subgroups and the full sample. Of greatest importance in this regard is whether the subgroup finding is statistically significantly different from that for the rest of the sample.

#### When Subgroup Estimates of Intervention Effects Differ Statistically Significantly

If the *difference* between a statistically significant estimated effect for a prespecified subgroup and that for other sample members is statistically significant, the subgroup finding can be considered confirmatory. However, a statistically significant difference in subgroup effects is usually difficult to demonstrate because of the limited power of statistical tests to identify such differences. 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 effect for their combined sample. Hence, seemingly large subgroup differences in estimated intervention effects often are not statistically significant. Ensuring that one's sample is large enough to detect such differences is thus an important benefit of prespecifying key subgroups while a study is being planned.

To provide an example of significant subgroup differences in intervention effects, we use results from a recent study of a program called Working toward Wellness (Kim, Leblanc,

and Michalopoulos, 2009). This program is being implemented in the Rhode Island site of MDRC's study of Enhanced Services for the Hard-to-Employ.[3] Parents (mostly mothers) receiving Medicaid in Rhode Island were recruited to participate in the study if they were judged to be depressed based on a series of questions in a screening interview.

Half of the study sample was then randomized to a program group, which received outreach from Master's-level clinicians who encouraged program group members to seek treatment for their depression. The other half of the sample was randomized to a control group, which did not receive this special encouragement but were eligible for all other services available to Medicaid recipients in Rhode Island. At six and 18 months after random assignment, sample members were interviewed and administered a set of questions to assess their level of depression. In addition, their medical claims data were obtained from the managed care organization that provides Medicaid services in Rhode Island to measure their receipt of services.

Before the impact analysis was conducted, but after the study was designed and the sample was enrolled, the research team decided to produce separate findings for Hispanic and non-Hispanic sample members because a prior study had found Hispanics to be easier to enroll in treatment for depression (Wells et al., 2000; Wells et al., 2004). Table 1 presents results for the full study sample and the two subgroups with respect to (1) a follow-up measure of the severity of depression based on a 30-point scale and (2) the percentage of sample members who were judged to be no longer depressed because they scored below 5 on the 30-point depression scale.

The estimated effects on both outcome measures were not statistically significant for the full study sample (denoted by the absence of "stars" for these findings). In addition, these estimates were not statistically significant for non-Hispanic sample members. However, they were statistically significant for Hispanic sample members. Furthermore, the findings for Hispanics were statistically significantly *different* from their counterparts for non-Hispanics (denoted by the presence of "daggers" for these findings). Thus it is appropriate to conclude that the program reduced depression among Hispanic sample members and to emphasize this finding when reporting results of the study. In addition, given the findings, the best existing estimates of program effects for Hispanics are an average reduction of 2.3 points on a 30-point scale of depression severity and a 13.7 percentage-point reduction in the likelihood of being depressed at follow-up. These findings are reported in the table as point estimates with corresponding levels of statistical significance.

[3]The study is being funded by the Administration for Children and Families and the Office of the Assistant Secretary for Planning and Evaluation of the U.S. Department of Health and Human Services.

**Table 1**

**Significant Differences Between Subgroup Findings**

**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 |
|---|---|---|---|---|
| **Depression severity (30-point scale)** | | | | |
| Full sample | 12.5 | 12.8 | -0.4 | 0.509 |
| Hispanic subgroup | 12.6 | 14.9 | -2.3 * | 0.049 † |
| Non-Hispanic subgroup | 12.4 | 12.0 | 0.4 | 0.531 † |
| **No longer depressed (%)** | | | | |
| Full sample | 12.3 | 9.9 | 2.4 | 0.463 |
| Hispanic subgroup | 11.7 | -2.0 | 13.7 ** | 0.005 †† |
| Non-Hispanic subgroup | 11.8 | 15.1 | -3.3 | 0.460 †† |

SOURCE: Findings were based on responses to a six-month follow-up survey.

NOTES: Statistical significance levels for the full sample and individual subgroups are indicated as: ***=0.1 percent,; ** = 1 percent; * = 5 percent. Statistically significant differences between the Hispanic and non-Hispanic subgroup are indicated as ††† = 0.1 percent; †† = 1 percent; † = 5 percent.

Another way to view these results is in terms of confidence intervals. Since the statistical question of interest is whether effects differ between the subgroups, the relevant confidence interval is for the difference in impacts. For the impact on depression severity between Hispanic and non-Hispanic sample members, the absolute difference in impacts on depression severity between Hispanic and non-Hispanic sample members in Rhode Island is 2.7 (-2.3 – 0.4), and the standard error of that absolute difference is 1.33 (not shown in the table). The 95 percent confidence interval ranges from 0.09 to 5.30. Although the full confidence interval is above zero — which is consistent with there being a significant difference between the subgroups — the confidence interval reveals considerable uncertainty regarding the true difference in effects between the subgroups.

### When Subgroup Estimates of Intervention Effects Do Not Differ Statistically Significantly

When statistically significant estimated effects for a subgroup are not statistically significantly different from those for the rest of their study sample, the next step in determining the confirmatory status of the subgroup estimate is to examine the statistical significance of estimated effects for the full sample and for the rest of the sample. In doing so, there are four possible cases to consider.

**Table 2**

**Significant Impacts for Both Subgroups and Full Sample**

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

| Outcome | Program Group | Control Group | Difference (Impact) | P-Value |
|---|---|---|---|---|
| **Employment in Year After Random Assignment (%)** | | | | |
| Full sample | 81.5 | 57.5 | 23.9 *** | <0.001 |
| Released from prison less than 3 months before entering study | 86.6 | 54.0 | 32.7 *** | <0.001 |
| Released from prison more than 3 months before entering study | 79.8 | 58.2 | 21.6 *** | <0.001 |

SOURCE: Based on data from state unemployment insurance records for sample members.

NOTES: Statistical significance levels for the full sample and individual subgroups are indicated as: *** = 0.1 percent; ** = 1 percent; * = 5 percent. The differences in impacts between the two subgroups were not statistically significant.

**Case 1: All impact estimates are statistically significant.** The simplest situation to interpret is when the estimated effect for the full study sample is statistically significant and all subgroup estimates are statistically significant and in the same direction. Findings in this case are confirmatory for all subgroups that were prespecified.

Table 2 illustrates this situation with results from MDRC's study of a transitional jobs program for male ex-offenders operated by the Center for Employment Opportunity (Redcross et al., 2009). The study focused on male residents of New York City who had been released from prison. Sample members who were randomized to the study's treatment group were offered a six-month subsidized job and were eligible for whatever other employment services were available to ex-offenders at the time. Sample members who were randomized to the study's control group were eligible only for the other existing employment services. The goal of the intervention was to provide a temporary economic base of support while treatment-group members sought permanent unsubsidized employment.

The literature on interventions to help ex-offenders reduce recidivism suggests that it is important to intervene as soon as possible after a person has been released from prison, because the risk of recidivism is highest during the time immediately after release (National Research Council, 2007). However, the CEO sample included many people who had been out of prison for a substantial period of time. To test the hypothesis that programs such as this are more effective for recently released prisoners than for others, researchers divided the sample roughly

**Table 3**

**Nonsignificant 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: Based on data from United Behavioral Health medical claims records.

NOTES: Statistical significance levels for the full sample and individual subgroups are indicated as: *** = 0.1 percent; ** = 1 percent; * = 5 percent. Differences in impacts between the Hispanic and non-Hispanic subgroups were not statistically significant.

in half into a subgroup that had entered the study sample within three months of release from prison and a subgroup that had been out of prison longer when they entered the sample.

The first step in the CEO study was to examine the extent to which providing subsidized jobs increased employment in the short term. Table 2 reports resulting estimates of the program's effects on sample members' employment during their first year after random assignment. These findings are based on data from state unemployment insurance records for individual sample members. The results indicate that for the full sample and for each of its two subgroups, the intervention increased postrelease employment rates substantially (by 23.9 percentage points for the full sample, 32.7 percentage points for recently released sample members, and 21.6 percentage points for other sample members). Researchers therefore concluded that CEO increased employment in the short term for both subgroups.

**Case 2: No impact estimates are statistically significant.** A second case can occur when estimated intervention effects are not statistically significant for a full study sample or for a set of its subgroups (for example, men versus women). In this case, the most that can be said about a subgroup in the set is that the study did not find convincing evidence of an intervention effect for it.

Table 3 illustrates this case using findings from the Rhode Island Working toward Wellness study introduced earlier. The outcome measure used for this example is the proportion of sample members who had filled prescriptions for antidepressants during the study's first 18 months of follow-up. This outcome measure, which was constructed from medical claims data for sample members, indicates small and non-statistically significant intervention effects for the full study sample and for its Hispanic and non-Hispanic subgroups (Kim et al., forthcoming).

**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) | P-value |
|---|---|---|---|---|
| **Incarcerated during first follow-up year (%)** | | | | |
| Full sample | 49.5 | 55.4 | -5.9 | 0.064 |
| Released from prison less than 3 months before entering study | 48.5 | 60.8 | -12.3 * | 0.020 |
| Released from prison more than 3 months before entering study | 50.4 | 53.6 | -3.2 | 0.457 |

SOURCE: Information from New York State criminal justice records.

NOTES: Statistical significance levels for the full sample and individual subgroups are indicated as: ***=0.1 percent; ** = 1 percent; * = 5 percent. Differences in estimated impacts between the two subgroups were not statistically significant.

The study team therefore concluded that the intervention did not affect use of antidepressants by the full sample or by either subgroup over an 18-month period.

**Case 3: Impact estimates are statistically significant for only one subgroup.** A third scenario can occur when impact estimates do not differ across subgroups, intervention effects are statistically significant for a subgroup of interest, but they are not statistically significant for the rest of the study sample or for the full sample. In this case, we recommend that, other things being equal, results for each subgroup should be considered exploratory.

The rationale for this recommendation is as follows. First, the estimated effect is not statistically significant for the full study sample. Hence, the most precise estimate that exists does not provide evidence that the intervention is effective. Second, the estimated effects for the two subgroups are not statistically significantly different from each other. Hence, there is not strong evidence that the statistically significant result for one subgroup is in fact different from the non-statistically significant result for the other subgroup. Consequently, the best information that exists for both subgroups is the full-sample finding.

Table 4 illustrates this case using results from the CEO program for its full sample and its two subgroups described above (recently released versus other former prisoners). The measure of recidivism used for this analysis is the proportion of sample members who were re-incarcerated during their first year after random assignment.

Point estimates for the full sample and both subgroups suggest that CEO might have reduced recidivism. However, these estimates are not statistically significant for the full sample or

**Table 5**

**Significant Impacts for One Subgroup and the 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 |
|---|---|---|---|---|
| **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: Based on claims records from United Behavioral Health.

NOTES: Statistical significance levels for the full sample and individual subgroups are indicated as: *** = 0.1 percent; ** = 1 percent; * = 5 percent. Differences in estimated impacts between the Hispanic and non-Hispanic subgroups were not statistically significant.

for sample members who had not been released recently. They are only statistically significant for sample members who had been released recently. Furthermore, the difference between impact estimates for the two subgroups is not statistically significant. Consequently, the findings do not provide convincing evidence that the intervention produced different effects for the two subgroups.

It therefore can be argued that the best existing evidence for both subgroups is that for the full sample. Consequently, although the estimated effect for recently released sample members is large (12.3 percentage points) and statistically significant, we recommend that it not be considered a confirmatory finding. Rather, we recommend that it be considered an encouraging exploratory finding that warrants an attempted replication by further research. In fact, this hypothesis is being tested further by MDRC's study of transitional jobs for reentering prisoners in several Midwestern states (Bloom, 2009).

**Case 4: Impact estimates are statistically significant for one subgroup and for the full study sample.** This case is probably the most controversial. It occurs when a finding for a 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 sample is not statistically significant and, in addition, findings for the two subgroups are not statistically significantly different from each other.

Table 5 illustrates this case using results from the Rhode Island Working toward Wellness study. The outcome measure for the example is the proportion of sample members who received mental health services during their first six months after random assignment. Estimates in the table indicate that the program increased the proportion of all sample members who

received mental health services by 10.5 percentage points, which is statistically significant. In addition, the findings indicate that the program increased the proportion of Hispanic sample members who received mental health services by 17.6 percentage points, which is statistically significant. However the estimated effect is only 5.4 percentage points for non-Hispanic sample members, and it is not statistically significant. Lastly, the difference between the two subgroup estimates is not statistically significant.

The dilemma here is whether to conclude that the program benefits Hispanic sample members while saying nothing about non-Hispanic sample members or instead to conclude that the program produces positive effects for both subgroups. The former conclusion could be drawn from the fact that the estimate for only one of the subgroups (Hispanics) is statistically significant. The latter conclusion could be drawn from the fact that the full-sample finding is statistically significant, and subgroup findings are not statistically significant different from each other. Here are the two positions stated more generally.

- **Position A:** The finding for the subgroup of interest (Hispanic sample members in the Working toward Wellness example) is confirmatory (assuming that the subgroup distinction was prespecified), because it is statistically significant in its own right and consistent with the best information that exists for that subgroup, absent direct information for it (the corresponding full-sample result). This finding does not imply that the study found no intervention effect for the rest of the study sample (non-Hispanic sample members in the Working toward Wellness 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 is positive indirect evidence from results for the full sample).

- **Position B:** The finding for the subgroup of interest (Hispanic sample members in the example) is exploratory, because there is no statistically significant difference between findings for the subgroup and the rest of the study sample. To advertise the significant finding for the subgroup of interest makes it look (by comparison) that 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 most likely to be particularly interested in making a conclusive statement about a given subgroup, regardless of how it compares with other subgroups. Hence, they would be willing to use the statistically significant finding for the subgroup and that for the

full sample as evidence that the intervention produced a positive effect for the subgroup of interest. Proponents of Position B are most likely to be interested in how findings for different subgroups compare with one another. Consequently, if differences among them are not statistically significant, they would choose to highlight the overall study finding.

## Contextual Considerations

Two additional factors — internal and external contextual considerations — should also influence how subgroup findings are reported and interpreted. Doing so acknowledges 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 when 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. For example, in the Working toward Wellness study there were a number of significant differences in impact estimates between Hispanic and non-Hispanic sample members during the first six months after random assignment, and that pattern gave the research team more confidence that there was a true difference in the early effects of the intervention being studied.

By external contextual considerations, we mean features of findings that are external to a given study. For example, other things being equal, 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 thought-out theory should receive more prominent attention than results that do not meet these conditions. Including either or both of these considerations as part of one's basis for interpreting a study's findings can broaden their contextual basis and thereby deepen the overall analysis.

## Multiple Hypothesis Testing

In closing, it is important to consider (albeit briefly) one further issue: the problem of distortions to statistical inferences that can occur when multiple related hypothesis tests are conducted. Such distortions reflect the fact that although the level of statistical significance for a single hypothesis test controls the conditional risk of making a Type I error for it, the conditional risk of making one or more Type I errors within a group of related hypothesis tests is much greater. And the more hypotheses that are tested, other things being equal, the greater this risk

is.[4] This "multiple testing" problem has been largely ignored by past intervention studies but is currently receiving a great deal of attention. In fact, it is probably the main reason that greater emphasis is now being placed on properly interpreting subgroup analyses.

Currently, there are four main approaches to minimizing the risk of inferential error due to multiple hypothesis testing. One approach, which is a core recommendation of the present paper, is to explicitly distinguish between exploratory and confirmatory findings. Doing so makes it possible to apply more stringent standards for testing multiple hypotheses in analyses that are meant to produce confirmatory findings in ways that limit the reduction of statistical power produced by these standards (by confining the multiplicity that is being accounted for to a smaller number of hypothesis tests).

A second approach, which we also endorse, is to minimize the number of confirmatory hypothesis tests conducted by a given study. Doing so reduces the multiplicity of hypothesis tests involved and thereby reduces the risk of inferential problems produced by multiplicity. Selection of this small number of confirmatory hypotheses should occur well before any analyses are conducted for a study and, if possible, during the development of its proposal or design paper. Not only can such an early decision-making process reduce the margin for multiplicity to create inferential problems, but the intellectual discipline that results from imposing its constraints can substantially improve the overall quality of the subsequent research.

A third approach to protecting against incorrect statistical inferences due to multiple hypothesis testing is to create an omnibus hypothesis test about the intervention's effects that considers all outcome measures and subgroups together. A popular version of this approach (Schochet, 2008) is to test the statistical significance of a composite measure of individual outcomes for all subgroups combined (that is, for the full study sample). If the estimated full-sample effect of the intervention on this composite outcome is statistically significant, this provides an additional source of confidence in the results of separate tests for individual outcome measures and subgroups. On the other hand, if the composite test does not indicate a statistically significant intervention effect, this result should be a source of skepticism about statistically significant findings for specific outcome measures or subgroups. Although this approach has some technical limitations (Schochet, 2008), we have used it in our research and would recommend it to others.

A fourth approach to guarding against incorrect statistical inferences produced by multiple hypothesis testing is to make adjustments (such as those named after Bonferonni [for example, Schochet, 2008, or Benjamini and Hochberg, 1995]) to the level of statistical signifi-

[4]By conditional risk we mean the risk or probability of obtaining an impact estimate that is statistically significant when the true impact is zero.

cance (p-value) for each individual hypothesis test in order to account for the number of tests conducted. These approaches overcompensate for multiple testing when outcome measures are correlated, and thereby unnecessarily reduce what is already limited statistical power for estimates of intervention effects. Even when they do not overcompensate, these approaches reduce statistical power considerably. For this reason, we have not used them in our research and are reluctant to recommend them to others.

## Conclusion

As noted, the purpose of this paper is to propose a set of criteria for deciding when and when not to highlight estimates of intervention effects for specific subgroups. These criteria are intended to reduce the likelihood of misinterpreting chance findings due to random error. Toward this end, we have recommended that researchers clearly distinguish between confirmatory and exploratory findings well in advance of conducting their analyses and that with rare exceptions they emphasize only subgroup findings that are statistically significant in their own right and differ statistically significantly from corresponding findings for other sample members. In addition, we have recommended that the validity of subgroup findings be judged in light of contextual considerations that are internal and external to the study being reported.

Although there is ultimately no way to guarantee that a given subgroup finding represents a true intervention effect, we believe that following a research protocol like that outlined above can greatly help to reduce the frequency with which chance findings are mistaken for true effects.

# References


Benjamini, Yoav, and Yosef Hochberg. 1995 "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society* 57, 1: 289-300.

Bloom, Dan. 2009. *The Joyce Foundation's Transitional Jobs Reentry Demonstration: Testing Strategies to Help Former Prisoners Find and Keep Jobs and Stay Out of Prison.* Chicago: The Joyce Foundation.

Bloom, Howard S., Saskia Levy Thompson, and Rebecca Unterman. 2010. *Transforming the High School Experience: How New York City's New Small Schools Are Boosting Student Achievement and Graduation Rates.* New York: MDRC.

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

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.

National Research Council. 2007. *Crime, Parole, Desistance from Crime, and Community Integration.* Committee on Community Supervision and Desistance from Crime. Washington, DC: The National Academies Press.

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: 176-186.

Schochet, Peter M. 2008. *Technical Methods Report: Guidelines for Multiple Testing in Impact Evaluations*." Washington, DC: U.S. Department of Education, Institute of Education Sciences.

Wells, K. B., Sherbourne, C., Schoenbaum, M., Duan, N., Meredith, L., Unutzer, J., Miranda, J., Carney, M. F., and Rubenstein, L. V. 2000. "Impact of Disseminating Quality Improvement Programs for Depression in Managed Primary Care: A Randomized Controlled Trial." *Journal of the American Medical Association* 283, 2: 212-220.

Wells, K., Sherbourne, C., Schoenbaum, M., Ettner, S., Duan, N., Miranda, J., Unutzer, J., and Rubenstein, L. 2004. "Five-Year Impact of Quality Improvement for Depression: Results of a Group-Level Randomized Controlled Trial." *Archives of General Psychiatry* 61, 4: 378-386.

# 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.