P/PV Brief
Public/Private Ventures

Evaluating Mentoring Programs

Jean Baldwin Grossman

September 2009

P/PV Public/Private Ventures is a national leader in creating and strengthening programs that improve lives in low-income communities. We do this in three ways:

INNOVATION
We work with leaders in the field to identify promising existing programs or develop new ones.

RESEARCH
We rigorously evaluate these programs to determine what is effective and what is not.

ACTION
We reproduce model programs in new locations, provide technical assistance where needed and inform policymakers and practitioners about what works.

P/PV is a 501(c)(3) nonprofit, nonpartisan organization with offices in Philadelphia, New York City and Oakland. For more information, please visit www.ppv.org.

Acknowledgments

This brief is a revised version of a chapter written for the *Handbook of Youth Mentoring* edited by David DuBois and Michael Karcher (2005). David and Michael provided many useful comments on the earlier version. Laura Johnson and Chelsea Farley of Public/Private Ventures helped revise the chapter to make it more accessible to non-mentoring specialists and provided great editorial advice.

Additional reference: DuBois, D. L. and M. J. Karcher, eds. 2005. *Handbook of Youth Mentoring*. Thousand Oaks, CA: Sage Publications, Inc.

Introduction

Questions about mentoring abound. Mentoring programs around the country are being asked by their funders and boards, "Does this mentoring program work?" Policymakers ask, "Does this particular type of mentoring—be it school-based or group or email—work?" These are questions about program impacts. Researchers and operators also want to know about the program's processes: What about mentoring makes it work? How long should a match last to be effective? How frequently should matches meet? Does the level of training, support or supervision of the match matter? Does parental involvement or communication matter? What types of interactions between youth and mentors lead to positive changes in the child? Then there are questions about the populations served and what practices are most effective: Are particular types of youth more affected by mentoring than others? Are mentors with specific characteristics, such as being older or more educated, more effective than other mentors or more effective with particular subgroups of youth? Finally, researchers in particular are interested in the theoretical underpinning of mentoring. For example, to what degree does mentoring work by changing children's beliefs about themselves (such as boosting self-esteem or self-efficacy), by shaping their values (such as their views about education and the future) or by improving their social and/or cognitive skills?

This article presents discussions of many issues that arise in answering both implementation or process questions and impact questions. Process questions are important to address even if a researcher is interested only in impacts, because one should not ask, "Does it work?" unless "it" actually occurred. The first section covers how one chooses appropriate process and impact measures. The next section discusses several impact design issues, including the inadequacies of simple pre/post designs, the importance of a good comparison group and several ways to construct comparison groups. The last section discusses common mistakes made when analyzing evaluation data and presents ways to avoid them. For a more complete discussion of evaluation in general, readers are referred to Rossi et al. (1999); Shadish et al. (2002); and Weiss (1998). Due to space constraints, issues entailed in answering mediational questions are not addressed here.

Measurement Issues

A useful guide in deciding what to measure is a program's logic model or theory of change: the set of hypothesized links between the program's action, participants' response and the desired outcomes. As Weiss states, with such a theory in hand, "The evaluation can trace the unfolding of the assumptions" (1998, 58). Rhodes et al. (2005) presents one possible theory of change for mentoring: Process measures describe the program's actions; outcome measures describe what effects the program has.

Process Measures

The first question when examining a program is: What exactly is the program as experienced by participants? The effect the program will have on participants depends on the realities of the program, not on its official description. All too frequently in mentoring programs, relatively few strong relationships form and matched pairs stop meeting. Process questions can be answered, however, at several levels. Most basically, one wants to know: Did the program recruit appropriate youth and adults? Did adults and youth meet as planned? Did all the components of the program happen? Were mentors trained and supervised as expected?

To address these questions, one examines the characteristics and experiences of the participants, mentors and the match, and compares them with the program's expectations. For example, a mentoring program targeting youth involved in criminal or violent activity tracked the number of arrests of new participants to determine whether they were serving their desired target populations (Branch 2002). A high school mentoring program for struggling students tracked the GPAs of enrolled youth (Grossman, Johnson 1999). Two match characteristics commonly examined are the average completed length of the relationship and the average frequency of interaction. Like all good process measures, they relate to the program's theory. To be affected, a participant must experience a sufficient dosage of the intervention. Some mentoring programs have more detailed ideas, such as wanting participants to experience specific program elements (academic support, for example, or peer interaction). If these are critical components of the program theory, they also make good candidates for process measures.

A second level of process question concerns the quality of the components: How good are the relationships? Are the training and supervision useful? These are more difficult dimensions to measure. Client satisfaction measures, such as how much youth like their mentors or how useful the mentors feel the training is, are one gauge of quality. However, clients' assessment of quality may not be accurate; as many teachers say, the most enjoyable class may not be the class that promotes the most learning. Testing mentors before and after training is an alternative quality measure. Assessing the quality of mentoring relationships is a relatively unexplored area. Grossman and Johnson (1999) and Rhodes et al. (2005) propose some measures.

From a program operator's or funder's perspective, how much process information is "enough" depends on striking a balance between knowing exactly what is happening in the program versus recognizing the service the staff could have provided in lieu of collecting data. Researchers should assess enough implementation data to be sure the program is actually delivering the services it purports to offer at a level and quality consistent with having a detectable impact before spending the time and money to collect data on outcomes. Even if no impact is expected, it is essential to know exactly what did or did not happen to the participants to understand one's findings. Thus, researchers may want to collect more process data than typically would be collected by operators to improve both the quality of their generalizations and their ability to link impacts to variation in participants' experiences of core elements of the program.

Lesson: Tracking process measures is important to program managers but essential for evaluators. Before embarking on an evaluation of impacts, be sure the program is delivering its services at a quality and intensity that would lead one to expect impacts.

Outcome Measures

An early task for an impact evaluator is to refine the "Does it work?" question into a set of testable evaluation questions. These questions need to specify a set of outcome variables that will be examined during the evaluation. There are two criteria for a good outcome measure (Rossi et al. 1999). First, the outcome can be realistically expected to change during the study period given the intensity of the intervention. Second, the outcome is measurable and the chosen measure sensitive enough to detect the likely change.

Evaluation questions are not program goals. Many programs rightly have lofty inspirational goals, such as enabling all participants to excel academically or to become self-sufficient, responsible citizens. However, a good evaluation outcome must be concrete, measurable and likely to change enough during the study period to be detected. Thus, for example, achieving a goal like "helping youth academically excel" could be gauged by examining students' grades or test scores.

In addition, when choosing the specific set of outcomes that will indicate a goal such as "academically excelling," one must consider which of the possible variables are likely to change given the program dosage participants will probably receive during the evaluation period. For example, researchers often have found that reading and math achievement test scores change less quickly than do reading or math grades, which, in turn, change less quickly than school effort. Thus, if one is evaluating the school-year (i.e., nine months) impact of a school-based mentoring program, one is likely to want to examine effort and grades rather than test scores, or at least in addition to test scores. Considerable care and thought need to go into deciding what outcomes data should be collected and when. Examining impacts on outcomes that are unlikely to change during the evaluation period can give the false impression that the program is a failure, when in fact the impacts on the chosen variables may not yet have emerged.

A good technique for selecting variables is to choose a range of proximal to more distal expected impacts based on the program's theory of change, which also represents a set of impacts ranging from modestly to impressively effective (Weiss 1998). Unfortunately, one cannot know a priori how long matches will last or how often the individuals will meet. Thus, it is wise to include some outcomes that are likely to change even with rather limited exposure to the intervention, and some outcomes that would change with greater exposure, thus setting multiple "bars." The most basic effectiveness goal is an outcome that everyone agrees should be achievable. From there, one can identify more ambitious outcomes.

Public/Private Ventures' evaluation of Big Brothers Big Sisters (BBBS) provides a good example of this process (Grossman and Tierney 1998). Researchers conducted a thorough review of BBBS's manual of standards and practices to understand the program's logic model and then, by working closely with staff from the national office and local agencies, generated multiple outcome bars. The national manual lists four "common" goals for a Little Brother or Little Sister: providing social, cultural and recreational enrichment; improving peer relationships; improving self-concept; and improving motivation, attitude and achievement related to schoolwork. Conversations with BBBS staff also suggested that having a Big Brother or Big Sister could reduce the incidence of antisocial behaviors such as drug and alcohol use and could improve a Little Brother's or Little Sister's relationship with his or her parent(s). Using previous research, the hypothesized impacts were ordered from proximal to distal as follows: increased opportunities for social and cultural enrichment, improved self-concept, better relationships with family and friends, improved academic outcomes and reduced antisocial behavior.

At a minimum, the mentoring experience was expected to enrich the cultural and social life of youth, even though many more impacts were anticipated. Because motivational psychology research shows that attitudes often change before behaviors, the next set of outcomes reflected attitudinal changes toward themselves and others. The “harder” academic and antisocial outcomes then were specified. Within these outcomes, researchers also hypothesized a range of impacts, from attitudinal variables, such as the child’s perceived sense of academic efficacy and value placed on education, to some intermediate behavioral changes, such as school attendance and being sent to the principal’s office, to changes in grades, drug and alcohol use, and fighting.

Once outcomes are identified, the next question is how to measure them. Two of the most important criteria for choosing a measure are whether the measure captures the exact facet of the outcome that the program is expected to affect and whether it is sensitive enough to pick up small changes. For example, an academically focused mentoring program that claims to increase the self-esteem of youth may help youth feel more academically competent but not improve their general feelings of self-worth. Thus, one would want to use a scale targeting academic self-worth or competence rather than a global self-worth scale—or select a scale that can measure both. The second consideration is the measure’s degree of sensitivity. Some measures are extremely good at sorting a population or identifying a subgroup in need of help but poor in detecting the small changes that typically result from programs. For example, in this author’s experience, the Rosenberg self-esteem scale (1979) is useful in distinguishing adolescents with high and low self-esteem but often is not sensitive enough to detect the small changes in self-esteem induced by most youth programs. On the other hand, measures of academic or social competency beliefs (Eccles et al. 1984) can detect relatively small changes.

Lesson: Choose outcomes that are integrally linked to the program’s theory of change, that establish multiple “effectiveness bars,” that are gauged with sensitive measures and that can be achieved within the evaluation’s time frame and in the context of the program’s implementation.

Choosing Informants

Another issue to be resolved for either process or outcome measures is from whom to collect information. For mentoring programs, the candidates are usually the youth, the mentor, a parent, teachers and school records. Information from each source has advantages and disadvantages. For example, for some variables, such as attitudes or beliefs, the youth may be the only individual who can provide valid information. Youth, for example, arguably are uniquely qualified to report on constructs such as their self-esteem (outcome measures) or considerations such as how much they like their mentors or whether they think their mentors support and care for them (process measures). Theoretically, what may be important is not what support the mentor actually gives but how supportive the youth perceives the mentor to be (DuBois et al. 2002).

On the other hand, youth-reported data may be biased. First, youth may be more likely to give socially desirable answers—recounting higher grades or less antisocial behavior. If this bias is different for mentored versus nonmentored youth, impact estimates based on these variables could be biased. Second, the feelings of youth toward their mentors may taint their reporting. For example, if the youth does not like the mentor’s style, he or she may selectively report or overreport certain negative experiences, such as the mentor missing meetings, and underreport others of a more positive nature, such as the amount of time the mentor spends providing help with schoolwork. Similarly, the youth may overstate a mentor’s performance to make the mentor look good. Last, the younger the child is, the less reliable or subtle the self-report. For this reason, when participants are quite young (8 or 9 years old), it is advisable to collect information from their parents and/or teachers.

The mentor often can be a good source of information about what the mentoring experience is like, such as what the mentor and mentee do and talk about (process measures), and as a reporter on the child’s behaviors at posttest (outcome measures). The main problem with mentor reporting is that mentors have an incentive to report positively on their relationships with youth and to see effects even if there are none, justifying why they are spending time with the child. Although there may be a positive bias, this does not preclude mentors’ being accurate in reporting relative impacts. This is because most mentors do not report that their mentees have improved equally in all areas. The pattern of difference in these reports, especially if consistent with those obtained from other sources, such as school records, may provide useful information about the true impacts.

Parents also can be useful as reporters. They may notice that the child is trying harder in school, for example, even though the child might not notice the change. However, like the mentor, parents may project changes that they wish were happening or be unaware of certain behaviors (e.g., substance use).

Finally, teachers may be good reporters on the behaviors of their students during the school day. Teachers who are familiar with age-appropriate behavior, for example, may spot a problem when a parent or mentor does not. However, teachers are extraordinarily busy, and it can be difficult for them to find the time to fill out evaluation forms on the participants. In addition, teachers too are not immune to seeing what they want to see, and as with mentors and parents, the caveat about relative impacts applies here.

Information also can be collected from records. Data about the occurrence of specific events—fights, cut classes, principal visits—are less susceptible to bias, unless the sources of these data (e.g., school administrators making discipline decisions) differentially judge or report events for mentored youth versus other youth.

Lesson: Each respondent has a unique point of view, but all are susceptible to reporting what they wish had happened. Thus, if time and money allow, it is advantageous to examine multiple perspectives on an outcome and triangulate on the impacts. What is important is to see a consistent pattern of impacts (not uniform consistency among the respondents). The more consistency there is, the more certain one can be that a particular impact occurred. For example, if the youth, parent and teacher data all indicate school improvement and test scores also increase, this would be particularly strong evidence of academic gains. Conversely, if only one of these measures exhibits change (e.g., parent reports), it could be just a spurious finding.

Design Issues

Answering the questions "Does mentoring work?" and "For whom?" may seem relatively straightforward—achievable simply by observing the changes in mentees' outcomes. But these ostensibly simple questions are harder to answer than one might assume.

The Fallacy of Pre/Post Comparisons

The changes we observe in the attitudes, behaviors or skills of youth while they are being mentored are not equivalent to program impacts. How can that be? The answer has to do with what statisticians call internal validity. Consider, again, the previously described BBBS evaluation. If one looks only at changes in outcomes for treatment youth (Grossman, Johnson 1999), one finds that 18 months after they applied to the program, 7 percent had reported starting to use drugs. On the face of it, it appears that the program was ineffective; however, during the same period, 11 percent of the controls had reported starting to use drugs. Thus, rather than being ineffective, this statistically significant difference indicates that BBBS was able to stem some of the naturally occurring increases in drug use.

The critical distinction here is the difference between outcomes and impacts. In evaluation, an outcome is the value of any variable measured after the intervention, such as grades. An impact is the difference between the outcome observed and what it would have been in the absence of the program (Rossi et al. 1999); in other words, it is the change in the outcome that was caused by the program. Simple changes in outcomes may in part reflect the program's impact but also might reflect other factors, such as changes due to maturation.

Lesson: A program's impact can be gauged accurately (i.e., be internally valid) only if one knows what would have happened to the participants had they not been in the program. This hypothetical state is called the "counterfactual." Because one cannot observe what the mentees would have done in the absence of the program, one must identify another group of youth, namely a comparison group, whose behavior will represent what the participants' behavior would have been without the program. Choosing a group whose behavior accurately depicts this hypothetical no-treatment (or counterfactual) state is the crux of getting the right answer to the effectiveness question, because a program's impacts are ascertained by comparing the behavior of the treatment or participant group with that of the selected comparison group.

Matched Comparison or Control Group Construction

There are two principal types of comparison groups: control groups generated through random assignment and matched comparison groups selected judgmentally by the researcher.

Experimental Control Groups

Random assignment is the best way to create two groups that would change comparably over time. In this type of evaluation, eligible individuals are assigned randomly, either to the control group and not allowed into the program, or to the treatment group, whose members are offered the program. (Note: "Treatments" and "controls" refer to randomly selected groups of individuals. Not all treatments may choose to participate. The term "participants" is used to refer to individuals who actually receive the program.)

The principal advantage of random assignment is that given large enough groups, on average, the two groups are statistically equivalent with respect to all characteristics, observed and unobserved, at the time the two groups are formed. If nothing were done to either group, their behaviors, on average, would continue to be statistically equivalent at any point in the future. Thus, if after the intervention the average behavior of the two groups differs, the difference can be confidently and causally linked to the program, which was the only systematic difference between the two groups. See Orr (1999) for a discussion of how large each group should be.

Although random assignment affords the most scientifically reliable way of creating two comparable groups, there are many issues that should be considered before using it. Two of the most difficult are "Can random assignment be inserted into the program's normal process without qualitatively changing the program?" and "Is it ethical to deny certain youth a mentor?" However, it is worth noting that all programs ration their services, primarily by not advertising to more people than they can serve. Random assignment gives all needy children an equal probability of being served, rather than denying children who need a mentor by not telling them about the program. The reader is referred to Dennis (1994) for a detailed discussion of the ethical issues involved in random assignment.

With respect to the first issue, consider first how the insertion of random assignment into the intake process affects the program. One of the misconceptions about random assignment among mentoring staff is that it means randomly pairing youth with adults. This is not the case. Random pairing would fundamentally change the program, and any evaluation of this altered program would not provide information on the effect of the actual program. A valid use of random assignment would entail randomly dividing eligible applicants between the treatment and control groups, then processing the treatment group youth just as they normally would be handled and matched. Under this design, random assignment affects only which youth files come across the staff's desk for matching, not what happens to youth once they are there. Another valid test would involve identifying two youth for every volunteer, then randomly assigning one child to the treatment group and one to the control group. For the BBBS evaluation, we used the former method because it was significantly less burdensome and emotionally more acceptable for the staff. However, the chosen design meant that not all treatment youth actually received a mentor. As will be discussed later, only about three quarters of the youth who were randomized into the treatment group and offered the program actually received mentors. (See Orr 1999 for a rich discussion of all aspects of random assignment.)

Matched (or Quasi-Experimental) Comparison Groups

Random assignment is not always possible. For example, programs may be too small or staff may refuse to participate in such an evaluation. When this is the case, researchers must identify a group of nonparticipant youth whose outcomes credibly represent what would have happened to the participants in the absence of the program. The weakness of the methodology is that the outcomes of the two groups can differ not only because one group got a mentor and the other did not but also because of other differences between the groups. To generate internally valid estimates of the program's impacts, one must control for the "other differences" either through statistical procedures such as regression analysis and/or through careful matching.

The researcher selects a comparison group of youth who are as similar as possible to the participant group across all the important characteristics that may influence outcomes in the counterfactual state (the hypothetical no-treatment state). Some key characteristics are relatively easy to identify and match for (e.g., age, race, gender or family structure). However, to improve the credibility of a matched comparison group, one needs to think deeply about other potential differences that could affect the outcome differential, such as whether one group of youth comes from families that care enough and are competent enough to search out services for their youth, or how comfortable the youth are with adults. These critical yet hard-to-measure variables are factors that are likely to systematically differ between participant and comparison group youth and to substantially affect one or more of the outcomes being examined. The more readers of an evaluation can think of such variables that have not been accounted for, the less they will believe the resulting program impact estimates.

Consider, for example, an email mentoring program. Not only would one want the comparison group to match the participant group on demographic characteristics—age (say, 12, 13, 14 or 15 years old), gender (male, female), race (white, Hispanic, black) and income (poor, nonpoor)—but one might also want to match the two groups on their preprogram use of the computer, such as the average number of hours per week spent using email or playing computer games. To match on this variable, however, one would have to collect computer use data on many nonparticipant youth to find those most comparable to the participants.

When one has more than a few matching variables, the number of cells becomes too numerous. In the above example, we would have 4 age × 2 gender × 3 race × 2 income, or 48 cells, even before splitting by computer use. A method that is used with increasing frequency to address this issue is propensity score matching (PSM). A propensity score is the probability of being a participant given a set of known factors. In simple random assignment evaluations, the propensity score of every sample member is 50 percent, regardless of his or her characteristics. In the real world, without random assignment, the probability of being a participant depends on the individual's characteristics, such as his or her comfort with computers in the example above. Thus, participants and nonparticipants naturally differ with regard to many characteristics. PSM can help researchers select which nonparticipants best match the participant group with respect to a weighted average of all these characteristics (where the weights reflect how important the factors are in making the individual a participant).

To calculate these weights, the researcher estimates, across both the participant and nonparticipant samples, a logistic model of the probability of being a participant (P_i) as a function of the matching variables and all other factors that are hypothesized to be related to participation (Rosenbaum, Rubin 1983; Rubin 1997). For example, if one were evaluating a school-based mentoring program, the equation might include age, gender, race, household status (HH) and reduced-price-lunch status (RL), as well as past academic (GPA) and behavior (BEH) assessments, as is shown in Equation 1 below. Obtaining teacher ratings of the youth's interpersonal skills (SOC) also would help match on the youth's ability to form a relationship.

$$(1)\ P_i = f(age,\ gender,\ race,\ HH,\ RL,\ GPA,\ BEH,\ SOC)$$

The next step of PSM is to compute for each potential member of the sample the probability of participation based on the matching characteristics in the regression. Predicted probabilities are calculated for both participants and all potential nonparticipants. Each participant then is matched with one or more nonparticipant youth based on these predicted propensity scores. For example, for each participant, the nonparticipant with the closest predicted participation probability can be selected into the comparison group. (See Shadish et al. 2002, 161–165, for further discussion of PSM, and Dynarski et al. 2003 for an application in a school-based setting.)

An implication of this technique is that one needs data for the propensity score logit from both the participant group and a large pool of nonparticipant youth who will be considered for inclusion in the comparison group. The larger the considered nonparticipant pool is, the more likely it is that one can find a close propensity score match for each participant. This data requirement often pushes researchers to select matching factors that are readily available through records rather than incur the expense of collecting new data.

One weakness of this method is that although the propensity to participate will be quite similar for the participant and comparison groups, the percentage with a particular characteristic (such as male) may not be, because PSM matches on a linear combination of characteristics, not each characteristic one by one. To overcome this weakness, most studies match propensity scores within a few demographically defined cells (such as race/gender).

PSM also balances the two groups only on the factors that went into the propensity score regression. For example, the PSM in Dynarski et al. (2003) was based on data gathered from 21,000 students to generate a comparison group for their approximately 2,500 participants. However, when data were collected later on parents, it turned out that comparison group students were from higher-income families. No matter how carefully a comparison group is constructed, one can never know for sure how similar this group is to the participant group on unmeasured characteristics, such as their ability to respond to adult guidance.

Lesson: How much a reader trusts the internal validity of an evaluation depends on how much he or she trusts that the comparison group truly is similar to the participant group on all important dimensions. This level of trust or confidence is quantifiable in random assignment designs (e.g., one is 95 percent confident that the two groups are statistically equivalent), whereas with a quasi-experimental design, this level of trust is uncertain and unquantifiable.

Analysis

This section covers how impact estimates are derived, from the simplest techniques to more statistically sophisticated ones. Several commonly committed errors and techniques used to overcome these problems are presented.

The Basics of Impact Estimates

Impact estimates for both experimental and quasi-experimental evaluation are basically determined by contrasting the outcomes of the participant or treatment group with those of the control or comparison group. If one has data from a random assignment design, the simplest unbiased impact estimate is the difference in mean follow-up (or posttest) outcomes for the treatment and control groups, as in Equation 2,

$$(2)\ b = \text{Mean}(Y_{fu,T}) - \text{Mean}(Y_{fu,C})$$

where b is the estimated impact of the program, $Y_{fu,T}$ is the value of outcome Y at posttest or follow-up for the treatment group youth, and $Y_{fu,C}$ is the value of outcome Y at posttest or follow-up for the control group youth. One can increase the precision of the impact estimate by calculating the change-score or difference-in-difference estimator as in Equation 3,

$$(3)\ b = \text{Mean}(Y_{fu,T} - Y_{bl,T}) - \text{Mean}(Y_{fu,C} - Y_{bl,C})$$

where $Y_{bl,T}$ is the value of outcome Y at baseline for the treatment group youth, and $Y_{bl,C}$ is the value of outcome Y at baseline for the control group youth.

Even more precision can be gained if the researcher controls for other covariate factors that affect the outcome through the use of regression, as in Equation 4,

$$(4)\ Y_{fu} = a + bT + cY_{bl} + dX + u$$

where b is the estimated impact of the program, T is a dummy variable equal to 1 for treatments and 0 for controls, and X is a vector of baseline covariates that affect Y and u (unmeasured factors). Another way to think of b is that it is basically the difference in the mean Ys, adjusting for differences in Xs.

When data are from a quasi-experimental evaluation, it is always best to estimate impacts using regression or analysis of covariance; not only does one get more precise estimates, but one can control for any differences that do arise between the participant and the comparison groups. Regression simulates what outcomes youth who were exactly like participants on all the included characteristics (the Xs) would have had if they had not received a mentor, assuming that all factors that jointly affect participation and outcomes are included in the regression. Regressions are also useful in randomized experiments for estimating impacts more precisely.

Suspicious Comparisons

The coefficient b from Equation 3 is an unbiased estimate of the program's impact (i.e., the estimate differs from the true impact only by a random error with mean of zero) as long as the two groups are identical on all characteristics (both included and excluded variables). The key to obtaining an unbiased estimate of the impact is to ensure that one compares groups of youth that are as similar as possible on all the important observable and unobservable characteristics that influence outcomes. Although many researchers understand the need for comparability and indeed think a lot about it when constructing a matched comparison group, this profound insight is often forgotten in the analysis phase, when the final comparisons are made. Most notably, if one omits youth from either group—the randomly selected treatment (or self-selected participant) group or the randomly selected control (or matched comparison) group—the resulting impact estimate is potentially biased. Following is a list of commonly seen yet flawed comparisons related to this concern.

Suspect Comparison 1: *Comparing groups of youth based on their match status, such as comparing those who received a mentor or youth whose matches lasted at least one month with the control or comparison group.* Suppose, as occurred in the Public/Private Ventures evaluation of BBBS's community-based mentoring program, only 75 percent of the treatment group actually received mentors (Grossman, Tierney 1998). Can one compare the outcomes of the 75 percent who were mentees with the controls to get an unbiased estimate of the program's impact? No. All the impact estimates must be based on comparisons between the entire treatment group and the entire control group to maintain the complete comparability of the two groups. (This estimate often is referred to as the impact of the "intent to treat.")

There are undoubtedly factors that are systematically different between youth who form mentoring relationships and those who do not. The latter youth may be more difficult temperamentally, or their families may have decided they really did not want mentors and withdrew from the program. If researchers remove these unmatched youth from the treatment group but do nothing with the control group, they could be comparing the "better" treatment youth with the "average" control group child, biasing the impact estimates. Randomization ensures that the treatment and control groups are equivalent (i.e., there are just as many "better" youth in the control group as the treatment group). After the intervention, matched youth are readily identified. Researchers, however, cannot identify the control group youth who would have been matched successfully had they been given the opportunity. Thus, if one discarded the unmatched treatment youth, implicitly one is comparing successfully matched youth to a mixed group—those for whom a match would have been found (had they been offered participation) and those for whom matches would not be found (who are perhaps harder to serve). An impact estimate based on such a comparison has the potential to bias the estimate in favor of the program's effectiveness. (The selection bias embedded in matching is the reason researchers might choose to compare the outcomes of a matched comparison group with the outcomes of mentoring program applicants, rather than participants.) On the other hand, the estimate based on all treatments and all controls, called the "intent-to-treat effect," is unaffected by this bias.

Because the intent-to-treat estimate is based on the outcomes of all of the treatment youth, whether or not they received the program, it may underestimate the "impact on the treated" (i.e., the effect of actually receiving the treatment). A common way to calculate the "impact on the treated" is to divide the intent-to-treat estimate by the proportion of youth actually receiving the program (Bloom 1984). The intent-to-treat estimate is a weighted average of the impact on the treated youth (a_p) and the impact on the untreated youth (a_{np}), as shown in Equation 5,

$$(5)\ \mathrm{Mean(T)} - \mathrm{Mean(C)} = a = p\, a_p + (1 - p)\ a_{np}$$

where p = proportion treated.

If the effect of group assignment on the untreated youth (a_{np}) is zero (i.e., untreated treatment individuals are neither hurt nor helped), then a is to equal a/p. Let's again take the example of the BBBS evaluation. Recall that 18 months after random assignment, 7 percent of the treatment group youth (the treated and untreated) had started using drugs, compared with 11 percent of the control group youth, a 4-percentage-point reduction. Using the knowledge that only 75 percent of the youth actually received mentors, the "impact on the treated" of starting to use drugs would increase from a 4-percentage-point reduction to a 5.3-percentage-point reduction (= 4/.75).

Similar bias occurs if one removes control group members from the comparison. Reconsider the school-based mentoring example described above, where treatment youth are offered mentors and control youth are denied mentors for one year. Suppose that although most youth participate for only a year, some continue their matches into a second school year. To gauge the impact of this longer intervention, the evaluator might (incorrectly) consider comparing youth who had mentors for two years with control youth who were not matched after their one-year denial period. This comparison has several problems. Youth who were able to sustain their relationships into a second year, for example, would likely be better able to relate to adults and perhaps more malleable to a mentoring intervention than the "average" originally matched comparison group member. An unbiased way to examine these program impacts would be to compare groups that were assigned randomly at the beginning of the evaluation: one group being offered the possibility of a two-year match and the other being denied the program for two years. To investigate both one- and two-year versions of the program, applicants would need to be randomized into one of three groups: one group offered the possibility of a two-year match, one group offered the possibility of a one-year match and one group denied the program for the full two years.

Lesson: The only absolutely unbiased estimate from a random assignment evaluation of a mentoring program is based on the comparison of all treatments and all controls, not just the matched treatments or those matched for *particular lengths of time.*

Suspect Comparison 2: *Comparing effects based on relationship characteristics, such as short matches with longer matches or closer relationships with less close relationships.* Grossman and Rhodes (2002) examined the effects of different lengths of matches using the BBBS evaluation data. In the first part of the paper, the researchers reported the straightforward comparisons between outcomes of those matched less than 6 months, 6 to 12 months and more than 12 months with the control group's outcomes. Although interesting, these simple comparisons ignore the potential differences among youth who are able to sustain their mentoring relationships for different periods of time. If the different match lengths were induced randomly across pairs or the reasons for a breakup were unrelated to the outcomes being examined, then there would be no problem with the simple set of comparisons. However, if, for example, youth who cannot form relationships that last more than five months are less able to get the adult attention and resources they need and consequently would do worse than longer-matched youth even without the intervention, then the first set of comparisons would produce biased impact estimates. Indeed, when the researchers statistically controlled for this potential bias (using two-staged least squares regression, as discussed below), they saw evidence of the strong association of short matches with negative outcomes disappear, while the indications of positive effects of longer matches remained.

A similar problem occurs when comparing youth with close relationships with those with weaker relationships. For the straightforward comparison to be valid, one is implicitly assuming that youth who ended up with close relationships with their mentors would have, in the absence of the program, fared equally well or poorly as youth who did not end up with close relationships. If those with closer relationships would have, without the program, been better able to secure adult attention than the other youth and done better because of it, for example, then a comparison of the close-relationship youth with either youth in less-close relationships or with the control/matched comparison group could be flawed.

Lesson: Any examination of groups defined by a program variable—such as having a mentor, the length of the relationship, having a cross-race match—is potentially plagued by selection bias regardless of the evaluation design employed. Valid subgroup estimates can be calculated only for subgroups defined on preprogram characteristics, such as gender or race or preprogram achievement levels or grades. In these cases, we can precisely identify and make comparisons to a comparable subgroup within the control group (against which the treatment subgroup may be compared).

Suspect Comparison 3: *Comparing the outcomes of mentored youth with a control or matched comparison group when the sample attrition at the follow-up assessment is substantial or, worse yet, when there is differential attrition between the two groups.* Once again, unless those who were assessed at posttest were just like the youth for whom one does not have posttest data, the impact estimates may be biased. Suppose youth from the most mobile, unstable households are the ones who could not be located. Comparing the "found" treatment and controls only provides information about the impact of the program on youth from stable homes, not all youth. This is an issue of generalizability (i.e., external validity; see Shadish et al. 2002).

Differential attrition between the treatment and the control (or participant and comparison) groups is important because it also poses a threat to internal validity. Frequently, researchers are able to reassess a much higher fraction of program participants—many of whom may still be meeting with their mentors—than of the control or comparison group youth (whom no one has necessarily tracked on a regular basis). For example, if the control or comparison group youth demonstrate increased behavioral or academic problems over the sample period, parents may move their children to attend other schools and thus make data collection more difficult. Alternatively, some treatment families may have decided not to move out of the area because the children had good mentors. Under any of these scenarios, comparing the reassessed comparison group youth with reassessed mentees could be a comparison of unlike individuals.

Technically, any amount of attrition—even if it is equal across the two groups—puts the accuracy of the impact estimates into question. The treatment group youth who cannot be located may be fundamentally different from control group youth who cannot be located. For example, the control attriters might be the youth whose parents enroll them in new schools because they are not doing well, while the treatment attriters might be the youth whose parents moved. However, as long as one can show that the baseline characteristics of the two groups are similar, most readers will accept the hypothesis that the two groups of follow-up responders are still similar. Similarly, if the baseline characteristics of the attriters are the same as those of the responders, then we can be more confident that the attrition was simply random and that the impact on the responders is indicative of the impact on all youth.

Lesson: Comparisons of treatment (or participant) groups and control (or comparison) groups are completely valid only if the youth not included in the comparison are simply a random sample of those included. This assumption is easier to believe if the nonincluded individuals represent a small proportion of the total sample, the baseline characteristics of nonresponders are similar to those of responders and the proportions excluded are the same for the treatment and control groups.

Statistical Corrections for Biases

What if one wants to examine program impacts under these compromised situations—such as dealing with differential attrition or examining the impact of mentoring on youth whose matches have lasted more than a year? There are a variety of statistical methods to handle these biases. As long as the assumptions underlying these methods hold, then the resulting adjusted impact estimates should be unbiased.

Let's start by restating the basic hypothesized model:

(6) $Y_{fu} = a + bM + cY_{bl} + dX + u$

The value of outcome Y_{fu} is determined by its value at baseline (Y_{bl}), whether the child got mentoring (M), a vector of baseline covariates that affect Y (X) and unmeasured factors (u). Suppose one has information on a group of mentees and a comparison group of youth matched on age, gender and school. Now suppose, however, the youth who actually get mentors differ from the comparison youth in that they are more likely to be firstborn. If firstborn youth do better on outcome Y (even controlling for the baseline level of Y) and one fails to control for this difference, the estimated impact coefficient (b) will be biased upward, picking up not only the effect of mentoring on Y but also the "firstborn-ness" of the mentees. The problem here is that M and u are correlated.

If one hypothesizes that the only way the participating youth differ from the average nonparticipating youth is on measurable characteristics (Z)—for example, they are more likely to be firstborn or to be Hispanic—then including these characteristics in the impact regression model, Equation 7, will fully remove the correlation between M and u, because M conditional on (i.e., controlling for) Z is not correlated with u. Thus, Equation 7 will produce an unbiased estimate of the impact (b):

(7) $Y_{fu} = a + bM + cY_{bl} + dX + fZ + u$

Including such extra covariates is a common technique. However if, as is usually the case, one suspects (or even could plausibly argue) that the mentored group is different in other ways that are correlated with outcomes and are unmeasured, such as being more socially competent or from better-parented families, then the estimate coefficient still will be potentially biased.

Instrumental Variables or Two-Staged Least Squares

Using instrumental variables (IV), also called two-staged least squares regression (TSLS), is a statistical way to obtain unbiased (or consistent) impact estimates in this more complicated position (see Stock and Watson 2003, Chapter 10).

Consider the factors influencing *M* (whether the child is a mentee)*:*

(8) $M = k + mZ + nX + v$

where *Z* represents variables related to *M* that are unrelated to *Y, X* represents variables related to *M* that are related to *Y* and *v* is the random error.

Substituting Equation 8 into Equation 6 results in:

(9) $Y_{fu} = a + b(k + mZ + nX + v) + cY_{bl} + u$

The problem is that *v* (the unmeasured elements related to participating in a mentoring program, such as having motivated parents) is correlated with *u.* This correlation will cause the regression to estimate a biased value for *b.* However, using instrumental variables, we are able to purge out *v* (the elements of *M* that are correlated with *u*) to get an unbiased estimate of the impact. Intuitively, this technique constructs a variable that is not *M* but is highly correlated with *M* and is not correlated with *u* (an "instrument").

The first and most difficult step in using this approach is to identify variables that 1) are related to why a child is in the group being examined, such as being a mentee or a long-matched child, and 2) are *not* related to the outcome *Y.* These are very hard to think of, must be measured for both treatment and control youth, and need to be considered before data collection starts. Examples might include the youth's interests, such as sports or outdoor activities, or how difficult it is for the mentor to drive to the child's home. These variables would be related to the match "working" (i.e., having longer duration) but not related theoretically to the child's grades or behaviors.

Then one estimates the following regression of *M:*

(10) $M = k + mZ + nX + cY_{bl} + w$

where *w* is a random error. All of the covariates that will be included in the final impact Equation 7, *X* and Y_{bl} are included in the first-stage regression along with the instruments *Z.* A predicted value of *M* ($M' = k + mZ + nX + cY_{bl}$) is then computed for each sample member. The properties of regression ensure that *M'* will be uncorrelated with the part of Y_{fu} not accounted for by Y_{bl}, or *X* (i.e., *u*). *M'* then is used in Equation 7 rather than *M.* The second stage of TSLS estimates Equation 7 and the corrected standard errors (see Stock and Watson 2003 for details). This technique works only if one has good predictive instruments. As a rule of thumb, the F-test for the Stage 1 regression should have a value of at least 10 if the instrument is to be considered valid.

Baseline Predictions

Suspect Comparison 2 illustrates how any examination of groups defined by a program variable, such as having a long relationship or a cross-race match, is potentially plagued by the type of selection bias we have been discussing. Schochet et al. (2001) employed a remarkably clever nonstatistical technique for estimating the unbiased impact of a program in such a case. The researchers knew they wanted to compare the impacts of participants who would choose different versions of a program. However, because one could not know who among the control group would have chosen each program version, it appeared that one could not make a valid comparison. To get around this problem, they asked the intake workers who interviewed all applicants before random assignment (both treatments and controls) to predict which version of the program each youth would end up in if all were offered the program. The researchers then estimated the impact of Version A (and similarly B) by comparing the outcomes of treatment and control group members deemed to be "A-likely" by the intake workers. Note that they were not comparing the treatment youth who actually did Version A to the A-likely control youth, but rather comparing the A-likely treatments to the A-likely controls. Because the intake workers were quite accurate in their predictions, this technique is convincing. For mentoring programs, staff could similarly predict which youth would likely end up receiving mentors or which would probably experience long-term matches based on the information they gathered during the intake process and their knowledge of the program. This baseline (preprogram) characteristic then could be used to identify a valid comparison.

Future Directions

Synthesis

Good evaluations gauge a program's impacts on a range of more to less ambitious outcomes that could realistically change over the period of observation given the likely program dosage; they assess outcomes using measures that are sensitive enough to detect the expected or policy-relevant change; and they use multiple measures and perspectives to assess an impact.

The crux of obtaining internally valid impact estimates is knowing what would have happened to the members of the treatment group had they not received mentors. Simple pre/post designs assume the participant would not have changed—that the postprogram behavior would have been exactly what the preprogram behavior was without the program. This is a particularly poor assumption for youth. Experimental and quasi-experimental evaluations are more valid because they use the behavior of the comparison group to represent what would have happened (the counterfactual state).

The internal validity of an evaluation depends critically on the comparability of the treatment (or participant) and control (or comparison) groups. If one can make a plausible case that the two groups differ on a factor that also affects the outcomes, the estimated impact may be biased by this factor. Because random assignment (with sufficiently large samples) creates two groups that are statistically equivalent in all observable and unobservable characteristics, evaluations with this design are, in principle, superior to matched comparison group designs; matched comparison groups can, at best, assure comparability only on the important observable characteristics.

Evaluators using matched comparison groups must always worry about potential selection-bias problems; in practice, researchers conducting random assignment evaluations often run into selection-bias problems too by making comparisons that undermine the balanced nature of treatment and control groups. Numerous statistical techniques, such as the use of instrumental variables, have been developed to help researchers estimate unbiased program impacts. However, their use requires forethought at the data collection stage to ensure that one has the data needed to make the required statistical adjustments.

Recommendations for Research

Given the aforementioned issues, researchers evaluating mentoring programs should consider the following suggestions:

1. *Design for disaster. Assume things will go wrong.* Random assignment will be undermined. There will be differential attrition. The comparison group will not be perfectly matched. To guard against these problems, researchers should think deeply about how the two groups might differ if any of these problems were to arise, then collect data at baseline that could be used for matching or making statistical adjustments. It is also useful to give forethought to which program subgroups will be examined and to collect variables that could help predict these program statuses, such as the length of a match.

2. *Gather implementation or process information.* This information is necessary to understand one's impact results—why the program had no effect or what type of program had the effects that were estimated. These data and data on program quality also can enable one to explore what about the program led to the change.

3. *Use random assignment or match on motivational factors.* Random assignment should be a researcher's first choice, but if quasi-experimental methods must be used, researchers should try to match participant and comparison youth on some of the less

obvious factors. The more one can convince readers that the groups are equivalent on all the relevant variables, including some of the hard-to-measure factors, such as motivation or comfort with adults, the more credible the impact estimates will be.

Recommendations for Practice

Given the complexities of computing valid impact estimates, what should a program do to measure effectiveness?

1. *Monitor key process variables or benchmarks.* Walker and Grossman (1999) argued that not every program should conduct a rigorous impact study: It is a poor use of resources, given the cost of research and the relative skills of staff. However, programs should use data to improve their programming (see United Way of America's *Measuring Program Outcomes* 1996 or the *W. K. Kellogg Foundation Evaluation Handbook* 2000*)*. Grossman and Johnson (1999) recommended that mentoring programs track three key dimensions: youth and volunteer characteristics, match length, and quality benchmarks. More specifically, programs could track basic information about youth and volunteers: what types and numbers apply, and what types and numbers are matched. They could also track information about how long matches last—for example, the proportion making it to various benchmarks. Last, they could measure and track benchmarks, such as the quality of the relationship (Rhodes et al. 2005). This approach allows programs to measure factors that (a) can be tracked easily and (b) can provide insight about their possible impacts without collecting data on the counterfactual state. Pre/post changes can be a benchmark (but not an impact estimate), and one must be careful that the types of youth served and the general environment are stable. If the pre/post changes for cohorts of youth improve over time, for example, but the program now is serving less needy youth, the change in this benchmark tells little about the effectiveness of the program (the counterfactual states for the early and later cohorts differ).

2. *Collaborate with local researchers to conduct impact studies periodically.* When program staff feel it is time to conduct a more rigorous impact study, they should consider collaborating with local researchers. Given the time, skills and complexity entailed in conducting impact research, trained researchers can complete the task much more efficiently. An outside evaluation also may be believed more readily. Researchers, furthermore, can become a resource for improving the program's ongoing monitoring system.

References

Bloom, H. S.
1984 "Accounting for No-Shows in Experimental Evaluation Designs." *Evaluation Review*, 8, 225–246.

Branch, A. Y.
2002 *Faith and Action: Implementation of the National Faith-Based Initiative for High-Risk Youth.* Philadelphia: Branch Associates and Public/Private Ventures.

Dennis, M. L.
1994 "Ethical and Practical Randomized Field Experiments." In J. S. Wholey, H. P. Hatry and K. E. Newcomer, eds., *Handbook of Practical Program Evaluation.* San Francisco: Jossey-Bass, 155–197.

DuBois, D. L., B. E. Holloway, J. C. Valentine and H. Cooper
2002 "Effectiveness of Mentoring Programs for Youth: A Meta-Analytic Review." *American Journal of Community Psychology*, 30, 157–197.

DuBois, D. L., H. A. Neville, G. R. Parra and A. O. Pugh-Lilly
2002 "Testing a New Model of Mentoring." In G. G. Noam, ed. in chief, and J. E. Rhodes, ed., *A Critical View of Youth Mentoring (New Directions for Youth Development: Theory, Research, and Practice*, No. 93, 21–57). San Francisco: Jossey-Bass.

DuBois, D. L. and M. J. Karcher, eds.
2005 *Handbook of Youth Mentoring.* Thousand Oaks, CA: Sage Publications, Inc.

Dynarski, M., C. Pistorino, M. Moore, T. Silva, J. Mullens, J. Deke et al.
2003 *When Schools Stay Open Late: The National Evaluation of the 21st Century Community Learning Centers Program.* Washington, DC: US Department of Education.

Eccles, J. S., C. Midgley and T. F. Adler
1984 "Grade-Related Changes in School Environment: Effects on Achievement Motivation." In J. G. Nicholls, ed., *The Development of Achievement Motivation.* Greenwich, CT: JAI Press, 285–331.

Grossman, J. B. and A. Johnson
1999 "Judging the Effectiveness of Mentoring Programs." In J. B. Grossman, ed., *Contemporary Issues in Mentoring.* Philadelphia: Public/Private Ventures, 24–47.

Grossman, J. B. and J. E. Rhodes
2002 "The Test of Time: Predictors and Effects of Duration in Youth Mentoring Programs." *American Journal of Community Psychology*, 30, 199–206.

Grossman, J. B. and J. P. Tierney
1998 "Does Mentoring Work? An Impact Study of the Big Brothers Big Sisters Program." *Evaluation Review*, 22, 403–426.

Orr, L. L.
1999 *Social Experiments: Evaluating Public Programs with Experimental Methods.* Thousand Oaks, CA: Sage.

Rhodes, J., R. Reddy, J. Roffman and J. Grossman
2005 "Promoting Successful Youth Mentoring Relationships: A Preliminary Screening Questionnaire." *Journal of Primary Prevention*, 147-167.

Rosenbaum, P. R. and D. B. Rubin
1983 "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika*, 70, 41–55.

Rosenberg, M.
1979 "Rosenberg Self-Esteem Scale." In K. Corcoran and J. Fischer (2000). *Measures for Clinical Practice: A Sourcebook* (3rd ed.). New York: Free Press, 610–611.

Rossi, P. H., H. E. Freeman and M. W. Lipsey
1999 *Evaluation: A Systematic Approach* (6th edition). Thousand Oaks, CA: Sage.

Rubin, D. B.
1997 "Estimating Causal Effects from Large Data Sets Using Propensity Scores." *Annals of Internal Medicine*, 127, 757–763.

Schochet, P., J. Burghardt and S. Glazerman
2001 *National Job Corps Study: The Impacts of Job Corps on Participants' Employment and Related Outcomes.* Princeton, NJ: Mathematica Policy Research.

Shadish, W. R., T. D. Cook and D. T. Campbell
2002 *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston: Houghton Mifflin.

Stock, J. H. and M. W. Watson
2003 *Introduction to Econometrics.* Boston: Addison-Wesley.

Tierney, J. P., J. B. Grossman and N. L. Resch
1995 *Making a Difference: An Impact Study of Big Brothers/Big Sisters.* Philadelphia: Public/Private Ventures.

United Way of America
1996 *Measuring Program Outcomes.* Arlington, VA: United Way of America.

Walker, G. and J. B. Grossman
1999 "Philanthropy and Outcomes: Dilemmas in the Quest for Accountability." In C. T. Clotfelter and T. Ehrlich, eds., *Philanthropy and the Nonprofit Sector in a Changing America.* Bloomington: Indiana University Press, 449–460.

Weiss, C. H.
1998 *Evaluation.* Upper Saddle River, NJ: Prentice Hall.

W. K. Kellogg Foundation
2000 *W.K. Kellogg Foundation Evaluation Handbook.* Battle Creek, MI: W. K. Kellogg Foundation.

Public/Private Ventures
2000 Market Street, Suite 600
Philadelphia, PA 19103
Tel: (215) 557-4400
Fax: (215) 557-4469

New York Office
The Chanin Building
122 East 42nd Street, 42nd Floor
New York, NY 10168
Tel: (212) 822-2400
Fax: (212) 949-0439

California Office
Lake Merritt Plaza, Suite 1550
1999 Harrison Street
Oakland, CA 94612
Tel: (510) 273-4600
Fax: (510) 273-4619

www.ppv.org