on the days the tests were administered, and the program in South Carolina calculated pregnancy rates for all adolescent women in the area over several years.

For purposes of truly evaluating program effectiveness, the window of follow-up often used is too short. Furthermore, it is difficult to determine whether the results of early follow-up are likely to overestimate or underestimate actual program effectiveness. For example, students who participate in a program may have lower levels of sexual activity, higher levels of contraceptive use and fewer pregnancies in the short run, but these differences may lessen later on as program participants “catch up” with students not participating in the program.

On the other hand, short-term differences or similarities in teenage pregnancy between program and nonprogram students may change as the adolescents get older. In comparison to the entire period when adolescents are at risk of unintended pregnancy, the period covered by most evaluations is relatively short. However, extending the period of follow-up beyond one or two years presents additional costs and challenges for evaluation research. For example, students who remain in the study through the end of follow-up may come from families that are stabler and better off than are those of students who drop out or change schools and are lost to follow-up.

Finally, the evaluation of pregnancy prevention programs is based on adolescents’ reports of their own behavior before and after the intervention. Thus, it is important to determine, to the degree possible, whether adolescents are reporting actual behavior or what the program has taught them is the desirable answer. In the evaluation of Postponing Sexual Involvement, the investigators checked information on sexual behavior, contraceptive use, pregnancies and STDs obtained through follow-up phone interviews against data taken from the adolescents’ medical records at Grady Memorial Hospital. They found discrepancies in only 1% of the cases.

Magnitude of Effects

The literature on the comparison of effect size across research studies describes several techniques for calculating effect size depending on the outcome measure and analytical methods used by each study.15 One of the most commonly used measures of effect size is the standardized mean difference between treatment and control groups in the outcome measure reported. The standardization process usually involves dividing the difference between treatment and control scores by the standard deviation of the outcome score. (The standard deviation may be the standard deviation of the control group’s outcome score or it may be a standard deviation pooling treatment and control group scores. The assumption behind this procedure is that the actual population standard deviation is similar across groups.) When comparing effect size across research studies, as is done in meta-analyses, one of the principal reasons for standardizing mean differences is to eliminate variation among study effects resulting from the use of different outcome measures (e.g., different scales used to measure achievement or behavioral or psychological traits) and thereby create a common metric that can be compared across studies.

In this review, however, we are comparing treatment and control groups using the same measures across all studies (proportions sexually active, proportions using contraceptives or proportions becoming pregnant). It is therefore analytically accurate and substantively preferable to use unstandardized mean differences in this context. By doing so, we can compare the effects of each program according to the actual proportions of adolescents who modified their behavior relative to the proportion of controls who did so.

Tables 2, 3 and 4 show the levels of sexual experience, contraceptive use and pregnancy experience in the treatment and control group before each program began or in its early stages (baseline) and the levels after the program had been running for some time (follow-up). The tables also present the amount of change that occurred between baseline and follow-up for each group. Finally, the tables show the unstandardized mean difference between the treatment and the control group in the change that occurred during the period studied and indicate whether this effect was statistically significant. Because the statistical significance of the difference between groups in the amount of change is a result of both the levels of change and the size of the study groups, a small absolute effect for one program may be statistically significant, while a larger effect for another program may not.

Delivering Sexual Initiation

In any cohort of adolescents, the proportion who have had sexual intercourse increases with the passage of time. The goal of programs is therefore to reduce the proportionate increase in the number of teenagers initiating sexual activity. To determine whether the programs we reviewed achieved such reductions, we compared the rate of increase among students who participated in a program with the rate among students with similar characteristics who did not.

For example, among all eighth graders (boys and girls) who participated in the Postponing Sexual Involvement program, 25% were sexually experienced at baseline (Table 2). At the end of the ninth grade, 43% of the program students were sexually experienced (an overall increase of 18 percentage points). In comparison, the proportion of sexually experienced stu-