This article describes a simple way to assess the statistical power of experimental designs. The approach presented is based on the concept of a minimum detectable effect, *which, intuitively, is the smallest true impact that an experiment has a good chance of detecting. The article illustrates how to compute minimum detectable effects and how to apply this concept to the assessment of alternative experimental designs. Applications to impact estimators for both continuous and binary outcome measures are considered.*

MINIMUM DETECTABLE EFFECTS

A Simple Way to Report the Statistical Power of Experimental Designs

HOWARD S. BLOOM
New York University

WHAT IS A MINIMUM DETECTABLE EFFECT?

The *minimum detectable effect* of an experiment is the smallest effect that, if true, has an X% chance of producing an impact estimate that is statistically significant at the Y level.

X is the statistical power of the experiment for an alternative hypothesis equal to the minimum detectable effect. Y is the level of statistical significance used to decide whether or not a true effect exists.

For example, consider the minimum detectable effect of an experimental study of a job training program defined as the true positive earnings effect for which the experiment has 80% power, using a one-sided hypothesis test at the .05 significance level. If the minimum detectable effect of hypothetical Design A for the experiment is $500, it has an 80% chance of producing a significant positive impact estimate (indicating that a positive impact exists) when the true impact is $500. If the minimum detectable effect of hypotheti-

EVALUATION REVIEW, Vol. 19 No. 5, October 1995 547-556

cal Design B for the experiment is $1,000, it has an 80% chance of identifying a true $1,000 program impact. Therefore, Design A has considerably greater statistical power than Design B.

Note that the minimum detectable effect of an experiment is measured in the original units of the impact of interest (dollars in the example). It is not standardized like the widely used effect size measure developed by Cohen (1977) as the basis for assessing statistical power. Cohen's effect size measure equals the program impact divided by the standard deviation of the corresponding outcome variable. This measure provides a very useful way to compare impacts across different outcome variables, different studies, and different program areas. Thus it is used frequently in meta-analyses as the basis for pooling impact findings (for example, see Glass, McGraw, and Smith 1981). When applied to a specific program, however, the minimum detectable effect defined above is more directly interpretable than the standardized effect size measure.

Furthermore, the minimum detectable effect is quite simple to compute. Figure 1 illustrates this point by comparing two normal (bell-shaped) sampling distributions for an experimental impact estimator based on a treatment and control group difference of means or difference of proportions. The sampling distribution on the left is centered on the null hypothesis of zero impact (vertical Line A). The sampling distribution on the right is centered on the minimum detectable effect (vertical Line C).

In this context, consider the statistical decision rule to be used for determining whether or not a true program effect exists. Assume that a one-sided hypothesis test will be used to determine whether the program produced a positive impact. Further assume that the .05 level will be used as the threshold for statistical significance. In this case, the critical region for the hypothesis test lies to the right of a point that is 1.65 standard errors above zero (vertical Line B). If the impact estimate is equal to or greater than this value, one would reject the null hypothesis of zero impact and accept the alternative hypothesis of a positive impact. If the impact estimate is below this cutoff, one would not accept the alternative hypothesis.

Now consider what will happen if the true program impact equals the minimum detectable effect. If the minimum detectable effect is defined as the true positive impact with 80% power given our decision rule, then 80% of its sampling distribution must lie above Line B. This implies that the minimum detectable effect is 0.84 standard errors above Line B.

Now put the two pieces together. If Line C is 0.84 standard errors above Line B and Line B is 1.65 standard errors above zero, then line C is 2.49 standard errors above zero. In other words, the minimum detectable effect equals 2.49 times the standard error of the impact estimate. This is true for

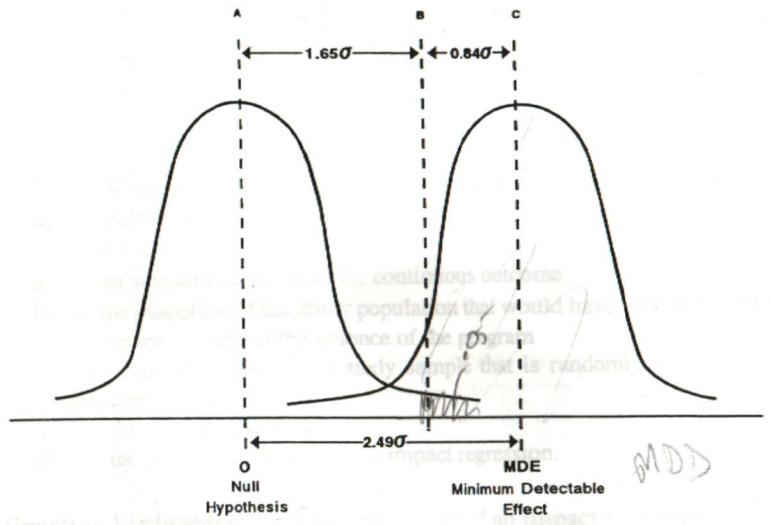

Figure 1: An Illustration of the Relationship Between the Minimum Detectable Effect and the Standard Error of an Impact Estimator

any minimum detectable effect with 80% power against a one-sided hypothesis test of zero impact at the .05 significance level.

For example, if the standard error of a program impact estimate is $500, the minimum detectable effect, as defined above, is 2.49 × $500, or $1,250. Thus $1,250 is the true program impact with an 80% chance of being identified (producing a significant positive impact estimate at the .05 level). True positive impacts larger than $1,250 will have more than an 80% chance of being identified. True positive impacts smaller than $1,250 will have less than an 80% chance of being identified.

This approach also applies to binary outcomes whose impacts can be expressed as proportions or percentages. For example, consider the impact of a program on high school graduation rates. If the standard error of the impact estimate is 0.05 (5 percentage points), the minimum detectable effect is 2.49 × 0.05, or 0.125 (12.5 percentage points). A true impact of this size would have an 80% chance of being identified. Larger true impacts would be more likely to be identified, and smaller true impacts would be less likely to be identified.

For any significance level, power value, and one- or two-sided hypothesis test, the minimum detectable effect can be computed as a multiple of the

TABLE 1: The Minimum Detectable Effect as a Multiple of the Standard Error of the Impact Estimate for Different Levels of Statistical Power and Statistical Significance

	Significance Level		
Statistical Power	*.10*	*.05*	*.01*
One-sided hypothesis test			
90%	256	2.93	3.61
80%	212	2.49	3.17
70%	180	2.17	2.85
Two-sided hypothesis test			
90%	293	3.24	3.86
80%	249	2.80	3.42
70%	217	2.48	3.10

standard error of the impact estimate. Table 1 lists this multiple for different definitions of the minimum detectable effect. The top panel presents the multiple for one-sided hypothesis tests. The bottom panel presents it for two-sided tests. The columns indicate the statistical significance level for the test, and the rows indicate its statistical power.

For example, the minimum detectable effect for a one-sided hypothesis test at the .05 level is 2.93 times the standard error at 90% power, 2.49 times the standard error at 80% power, and 2.17 times the standard error at 70% power.

HOW TO ESTIMATE A MINIMUM DETECTABLE EFFECT

To estimate the minimum detectable effect of a proposed experimental design, one must estimate the standard error of its impact estimate and take an appropriate multiple. Equation (1) indicates how to compute the standard error of an experimental impact estimate for a continuous outcome obtained from a regression-adjusted treatment/control group difference of means.[1] Equation (2) presents the corresponding expression for a binary outcome expressed as a proportion.[2]

$$\sigma_C = \sqrt{\frac{\sigma^2 (1 - R^2)}{T (1 - T) n}} \qquad (1)$$

$$\sigma_B = \sqrt{\frac{\Pi (1 - \Pi) (1 - R^2)}{T (1 - T) n}} \qquad (2)$$

where

σ_C = the standard error of the impact estimator for a continuous outcome
σ_B = the standard error of the impact estimator for a binary outcome, expressed as a proportion
σ = the standard deviation of the continuous outcome
Π = the proportion of the study population that would have a value of 1 for the binary outcome in the absence of the program
T = the proportion of the study sample that is randomly assigned to the treatment group
n = the size of the study sample
R^2 = the explanatory power of the impact regression.

Equation 1 indicates that the standard error of an impact estimator

- increases as the heterogeneity of the study population, σ^2, increases
- decreases as the explanatory power of the impact regression, R^2, increases
- decreases as the sample size, n, increases.

The equation also indicates that the standard error of an impact estimator depends on the proportion of the study sample assigned to the treatment group, T, an issue to which we return later. Equation (2) differs from Equation (1) only in that the population variance is expressed as $\Pi(1 - \Pi)$ for a binary outcome instead of σ^2 for a continuous outcome.[3]

To estimate the minimum detectable effect of a treatment/control group comparison, one must assume values for σ or Π and R^2. To the extent possible, these assumptions should be based on previous research. It is then a simple matter to compute the standard error for any given sample size and treatment group assignment fraction. Having computed the standard error, one can then compute the minimum detectable effect by taking the multiple of the standard error that corresponds to the statistical power, statistical significance, and one-sided or two-sided hypothesis test to be used for the impact analysis.

Consider the following examples, which determine the minimum detectable effect with 80% power for a one-sided hypothesis test at the .10 significance level. In these cases, the minimum detectable effect is 2.12 times the standard error.[4]

Example 1

The minimum detectable effect of a job training program on the annual earnings of treatment group members, where

$\sigma = \$7{,}000$ $n = 500$
$R^2 = .20$ $T = 0.5$

Example 2

The minimum detectable effect of an educational program on the test scores of treatment group members, using scale scores for the Test of Adult Basic Education (TABE), where

$\sigma = 40$ scale points $n = 500$
$R^2 = .40$ $T = 0.5$

Example 3

The minimum detectable effect of a correctional program on the proportion of treatment group members who commit future crimes (the recidivism rate), where

$\Pi = 0.7$ $n = 500$
$R^2 = .05$ $T = 0.5$

Substituting the information for Example (1) into Equation (1) yields a standard error for the impact estimator of $560. Substituting the information for Example 2 into Equation (1) yields a standard error for the impact estimator of 2.8 scale points. Substituting the information for Example 3 into Equation (2) yields a standard error for the impact estimator of 0.040 (4.0 percentage points). Multiplying each standard error by 2.12 yields corresponding minimum detectable effects of $1,190 for Example 1, 5.9 scale points for Example 2, and 0.085 (8.5 percentage points) for Example 3.

Therefore, if the true program impact were +$1,190 for Example 1, +5.9 scale points for Example 2, and +8.5 percentage points for Example 3, each experiment would have an 80% chance of indicating that a positive program impact existed.[5]

TABLE 2: The Effect of the Treatment/Control Group Mix on the Minimum Detectable Effect (MDE) for Three Hypothetical Outcomes

Treatment/Control Group Mix	MDE for Example 1 (dollars)	MDE for Example 2 (scale points)	MDE for Example 3 (percentage points)	Ratio of MDE to Optimal MDE
50/50 (optimal)	1,190	5.9	8.5	1.00
60/40	1,210	6.0	8.7	1.02
70/30	1,300	6.4	9.3	1.09
80/20	1,490	7.3	10.6	1.25
90/10	1,980	9.8	14.2	1.67

HOW THE TREATMENT\CONTROL GROUP ALLOCATION AFFECTS THE MINIMUM DETECTABLE EFFECT

To see how one can use the minimum detectable effect to compare alternative experimental designs, consider the question of how to allocate one's study sample to treatment and control status.[6] This issue played a major role in the design of a recent large-scale experiment (Bloom et al. 1993). In that study, there was a conflict between the desire of program staff to minimize the number of sample members assigned to the control group and the desire of evaluators to maximize statistical power (which requires a 50/50 treatment/control group mix). To examine the tradeoffs involved, the evaluators computed minimum detectable effects for a range of different treatment/control group mixes.

Table 2 presents a similar analysis for the three examples discussed above. Column 1 in the table indicates the treatment\control group mix.[7] Columns 2 through 4 list the corresponding minimum detectable effects. The minimum detectable effects in the table for a 50/50 treatment/control group mix are the same as those computed above for each example.

To compute the minimum detectable effects for a 60/40 treatment/control group mix, one need only change the value of T in Equations (1) and (2) and hold the other parameters for each example constant. Note that there is almost no difference between the minimum detectable effect for a 60/40 mix and that for a 50/50 mix. In addition, continuing down each column, it can be seen that the minimum detectable effect increases quite slowly as the treatment/control group mix departs from optimality (50/50).[8] Hence one can deviate substantially from optimality in this regard and still maintain most of the statistical power of an experiment.

Column 5 in the table illustrates this point by presenting the ratio of the minimum detectable effect for each treatment/control group mix to the minimum detectable effect for the optimum 50/50 mix. For example, the minimum detectable effect for a 60/40 mix is 1.02 times that for a 50/50 mix, the minimum detectable effect for a 70/30 mix is 1.09 times that for a 50/50 mix, and so on.[9]

In light of these findings, evaluators in the study cited above chose a two-thirds/one-third treatment/control group mix, which greatly increased the political feasibility of their study without markedly reducing its statistical power.

STATISTICAL POWER AND ONE-SIDED VERSUS TWO-SIDED HYPOTHESIS TESTS

The main goal of a program impact study should be to determine whether or not a program produced the results it was intended to produce. For example, drug treatment programs are intended to reduce drug abuse, medical treatments are intended to improve health status, housing programs are intended to improve housing quality, correctional programs are intended to reduce crime, and so on. If such programs produce their intended effects at an acceptable cost, then they should be continued or expanded. If they produce no effect or an effect that is contrary to that intended, then the programs should be modified or eliminated.

One, therefore, should use a one-sided hypothesis test to determine the statistical significance of an impact estimate in this decision-making context. The null hypothesis should be no effect and the alternative hypothesis should be the intended effect. If a program is intended to increase an outcome (for example, access to health care), then one should test for a statistically significant increase in this outcome. If a program is intended to reduce an outcome (for example, smoking), then one should test for a statistically significant reduction in this outcome.

A one-sided hypothesis test has a smaller minimum detectable effect (greater statistical power) than a two-sided test. For example, if the minimum detectable effect at 80% power and .10 significance is $1,000 for a one-sided test, it is $1,175 for a two-sided test.

Unfortunately, the convention followed by most evaluators is to use a two-sided test unless the theory underlying the intervention of interest is strong enough to rule out the unintended effect. This convention comes from empirical social science research, in which the primary objective is to test theories about relationships between variables. For testing such theories, it has been judged to be most appropriate to use a two-sided hypothesis test of whether some relationship (direction unspecified) or no relationship exists.

Because program evaluations should be conducted to inform program decisions, evaluators should use the logic of the decision they are attempting to inform rather than the strength of existing theory to formulate their impact hypothesis test. This issue has important practical consequences, because limited statistical power is a major weakness of most program impact studies (Lipsey 1990). There are well-established ways to increase power, such as increasing sample size, increasing the reliability of outcome measures, blocking the experimental sample, and statistically controlling for relevant covariates. But shifting from a two-sided hypothesis test to a more powerful one-sided test is not usually prescribed, although it should be in many cases.

NOTES

1. Equations (1) and (2) apply when the experimental sample is a simple random sample from the study population. These expressions do not account for the *design effect* on the standard error produced by sampling sites in multisite experiments. Excellent discussions of this design effect are provided by Corson et al. (1994) and by Greenberg, Meyer, and Wiseman (1992).

2. A standard regression-adjusted difference of means for estimating program impacts on a continuous outcome includes a dummy variable for treatment/control status plus covariates representing baseline characteristics of sample members. The estimated regression coefficient for the treatment/control dummy variable is the program impact estimate. One can also specify a binary (0, 1) outcome as the dependent variable of such a regression. The resulting linear probability model provides impact estimates that reflect a regression-adjusted treatment/control group difference of proportions. More complex logit or probit models for binary dependent variables generally produce hypothesis test findings that are similar to those from linear probability models, unless the underlying probabilities are near 0 or near 1. Because statistical power analyses usually are based on *assumed* parameter values, the approximation provided by a linear probability model usually will be justified by its simplicity.

3. The standard error of a nonexperimental impact estimator obtained from a regression-adjusted difference of means or difference of proportions for a treatment group and a nonrandom comparison group has an additional term in the denominator of Equations (1) and (2). This term accounts for any multicollinearity that exists between the treatment/comparison group dummy variable and the covariates included in the impact regression. It equals 1 minus the R^2 obtained from regressing the treatment/control group dummy variable on all of the covariates in the impact model.

4. The parameters used for the three examples are similar to those found by past studies. Hence the minimum detectable effects presented in the text reflect realistic values.

5. Scale scores for the TABE have a nonlinear relationship to grade equivalent scores. However, for the eighth-grade range, 1 point on the scale score is roughly equivalent to a 0.1-year grade equivalent. Therefore, a +5.9 minimum detectable effect in terms of scale scores reflects an increase of 0.6 years in grade-equivalent terms.

6. Lipsey (1990) and Kraemer and Thiemann (1987) formulate this issue differently but illustrate the same point.

7. Parallel findings for designs in which the control group is larger than the treatment group can be found in the table by reversing the order of the treatment and control group percentages. For example, the minimum detectable effect of a 40/60 treatment/control group mix is the same as that for a 60/40 mix, and so on.

8. This is because the minimum detectable effect increases in inverse proportion to the square root of $T(1 - T)$.

9. A ratio of 1.02 indicates that the minimum detectable effect for a 60/40 mix is 2% larger than that for a 50/50 mix. A ratio of 1.09 indicates that the minimum detectable effect for a 70/30 mix is 9% larger than that for a 50/50 mix, and so on.

REFERENCES

Bloom, Howard S., Larry L. Orr, George Cave, Stephen H. Bell, and Fred Doolittle. 1993. *The national JTPA study: Title IIA impacts on earnings and employment at 18 months.* Bethesda, MD: Abt Associates.

Bloom, Howard S., Larry L. Orr, Fred Doolittle, Joseph Hotz, and Burt Barnow. 1990. *Design of the national JTPA study.* Bethesda, MD: Abt Associates.

Cohen, Jacob. 1977. *Statistical power analysis for the behavioral sciences.* New York: Academic Press.

Corson, Walter S., Paul T. Decker, Terry R. Johnson, and Daniel H. Klepinger. 1994. *Job search assistance demonstration design report.* Princeton, NJ: Mathematica Policy Research.

Glass, Gene V., Barry McGraw, and Mary Lee Smith. 1981. *Meta-analysis in social research.* Beverly Hills, CA: Sage.

Greenberg, David, Robert H. Meyer, and Michael Wiseman. 1992. Prying the lid from the black box: Plotting evaluation strategy for employment and training programs. Paper presented at the Fourteenth Annual Research Conference of the Association for Public Policy Analysis and Management, 29-31 October, Denver.

Kraemer, Helena Chmura, and Sue Thiemann. 1987. *How many subjects? Statistical power analysis in research.* Newbury Park, CA: Sage.

Lipsey, Mark W. 1990. *Design sensitivity: Statistical power for experimental research.* Newbury Park, CA: Sage.

Metcalf, Charles. 1974. Alternative approaches to optimal sample assignment in the supported work evaluation. Working paper no. E-6, Mathematica Policy Research, Princeton, NJ.

Rosenthal, R., and D. B. Rubin. 1982. A simple, general purpose display of the magnitude of experimental effect. *Journal of Educational Psychology* 74:166-9.

Sechrest, L., and W. H. Yeaton. 1982. Magnitudes of experimental effects in social science research. *Evaluation Review* 6:579-600.

Woodbury, Stephen A., and Robert G. Spiegelman. 1987. Bonuses to workers and employers to reduce unemployment: Randomized trials in Illinois. *American Economic Review* 77 (4): 513-30.

Howard S. Bloom is the director of doctoral studies at the Robert F. Wagner Graduate School of Public Service, New York University. He is a specialist in evaluation research methodology and has extensive experience in the design, implementation, and analysis of large-scale randomized experiments to study the impacts of employment and training programs.

APPLICATION OF DIFFERENCE-IN-DIFFERENCE TECHNIQUES TO THE EVALUATION OF DROUGHT-TAINTED WATER CONSERVATION PROGRAMS

ANIL BAMEZAI