A cutoff-based randomized clinical trial couples cutoff-based assignment on an appropriate covariate with random assignment to help balance ethical and scientific concerns in certain situations. A statistical power algorithm based on the Fisher Z method is developed that is particular to and inclusive of cutoff-based random clinical trials and the single cutoff-point (regression-discontinuity) design, which has no randomization. This article quantifies power and sample size estimates for varying levels of randomization and cutoff-based assignment. Although more randomization engenders greater statistical power, less randomization requires a much larger increase in sample size for small treatment effects.

POWER ANALYSIS OF CUTOFF-BASED RANDOMIZED CLINICAL TRIALS

JOSEPH C. CAPPELLERI
New England Medical Center

RICHARD B. DARLINGTON
WILLIAM M.K. TROCHIM
Cornell University

In principle, the conventional or traditional randomized clinical trial (RCT) provides the most statistically powerful and scientifically rigorous method for comparing the efficacy of treatments. In recent years, however, ethical concerns about the RCT have been raised when strong a priori information exists that the test treatment is more beneficial than the control treatment. Trochim and Cappelleri (1992) presented a class of cutoff-based RCTs that combines randomized assignment with cutoff-based assignment to help balance scientific and ethical concerns in certain situations.

In a simple version of a cutoff-based RCT, subjects scoring below a cutoff score on a baseline measure (i.e., the least severely ill) are automatically assigned to the control-treated group, those scoring above a second, higher cutoff (i.e., the most ill) are automatically assigned to the test-treated group,

AUTHORS' NOTE: *This work was partially supported by grant number HS-07782-01 awarded by the Agency of Health Care Policy and Research (Dr. Joseph Lau, principal investigator). The authors thank Louise M. Ryan, an associate professor of biostatistics at the Harvard School of Public Health, for her helpful critique of the manuscript.*

EVALUATION REVIEW, Vol. 18 No. 2, April 1994 141-152

and those scoring in the interval between the cutoff scores (i.e., the moderately ill) are randomly assigned to either group. Trochim and Cappelleri (1992) also considered a single cutoff-point design with no randomization—known as the regression-discontinuity (RD) design—whereby all subjects scoring above a single cutoff value are automatically placed in one treatment group, whereas all subjects scoring below the same cutoff value are automatically placed in the other group. When properly modeled, both the cutoff-based RCT and the RD design give an unbiased effect of treatment (Berk and Rauma 1983; Cappelleri et al. 1991; Trochim, Cappelleri, and Reichardt 1991; Trochim and Cappelleri 1992; Reichardt, Trochim, and Cappelleri forthcoming).

More randomization engenders greater statistical power. Everything else the same, a cutoff-based RCT with less randomization has lower power to detect a treatment effect than a cutoff-based RCT with more randomization. Moreover, statistical power is lower in a cutoff-based RCT than in a traditional RCT. The question becomes "How many more subjects are needed in a cutoff-based RCT to reach the same level of power as the conventional RCT?"

Assuming a normally distributed baseline assignment variable with a cutoff point at the mean, Goldberger (1972), who was most influential in increasing our understanding on the efficiency and power of the cutoff-based designs, stated that the RD design requires about 2.75 times more cases than the RCT. He obtained this number by calculating the efficiency of the RCT relative to the RD design, using the ratio of variances of the regression coefficients for treatment, which is the reciprocal of one minus the square of the correlation between treatment and baseline variables. Taking into account a given treatment effect size and significance level, Cappelleri (1991) quantified the statistical power and sample sizes needed for cutoff-based designs with varying amounts of randomization and compared these results with the conventional RCT.

Several approaches are available to conduct power analysis for correlation and regression. One of the most popular in the behavioral and health sciences is the approach in Cohen (1988), which is based on the noncentral chi-square distribution. His method for multiple regression and correlation can be applied to cutoff-based designs. However, in this article we construct our own formulations that are specific to cutoff-based RCTs. Based on the Fisher Z method (Darlington 1990), an algorithm and its program have been developed that are particular to and inclusive of cutoff-based RCTs as well as the RD design, and a general methodology was developed to compare their power and sample size estimates to the conventional RCT (R. B. Darlington, personal communication, July 20, 1990; Cappelleri 1991).

THE CUTOFF-BASED POWER ALGORITHM

ASSUMPTIONS

Six basic restrictions were placed on the algorithm that generated sample size and power estimates. First, to simplify matters, we assumed that the baseline assignment variable X—on which treatment assignment is based in cutoff-based designs—was normally distributed. This was also done to generalize across baseline scales with different units of measurement. Second, the algorithm assumed that X was the only regressor other than the dichotomous treatment variable Z (Z = 1 if test treated, Z = 0 if control treated).

Third, the continuous outcome variable Y was formulated as

$$Y = X + B_z Z + e,$$

where Y is the continuous outcome variable, B_z is the treatment effect, and e is the residual term. Outcome and baseline variables do not necessarily have to belong to the same measure, and a transformation can occur on the original values of each variable.

Fourth, because it was presumed that higher scores on the baseline measure indicated a higher degree of illness, in the cutoff-based RCTs (and RD design), higher baseline scores favored assignment to test treatment, which was thought to be potentially more beneficial than control treatment. The exact results would be obtained if lower scores indicated a higher degree of illness. Fifth, the algorithm applied specifically to a one-sided test at the .025 significance level for the null hypothesis of nonassociation, although it tends to closely approximate the power of a two-sided test at the .05 level. Finally, the overall proportion of cases in the two groups was .50. For the cutoff-based RCTs, the proportion of cases within the interval of randomization assigned to either treatment was also .50.

The power-analytic procedure can be modified to include baseline scores following a given nonnormal distribution, additional covariates, and overall and within-interval proportions different from .50.

THE FISHER Z METHOD

We defined PR(Y, Z) to symbolize the population partial correlation between Y and Z (given X). The quantity $PR^2(Y, Z)$ can be thought of as the unique Y variance explained by Z, expressed as a proportion of the Y variance unexplained by X. As defined here, power is the probability of rejecting the null hypothesis that the true partial correlation between the outcome and treatment variables equals zero if it in fact equals some prespecified alternative. Therefore, H_0: PR(Y, Z) = 0 and H_1: PR(Y, Z) = pr, where pr is the alternative value taken as the true partial correlation. Testing this null hypothesis is tantamount to testing the null hypothesis of no treatment effect because the partial correlation between outcome Y and treatment Z equals zero if and only if the partial regression coefficient for treatment on outcome equals zero. The procedure has the flexibility of being applied to tests of null hypotheses other than nonassociation.

In the power analytic framework adopted, power and sample size estimates depended on three factors. One factor is Fisher's Z transformation, abbreviated as fz. We used the formula,

$$\text{fz} = .5 \ln \frac{[1 + \text{pr}]}{[1 - \text{pr}]} .$$

A second factor is the standard error of fz, denoted as se(fz), which for our purposes can be written as

$$\text{se(fz)} = \frac{1}{\sqrt{\{\text{number of cases}\} - 4}} .$$

A third factor is the z value corresponding to a specified level of significance and the direction of the alternative hypothesis.

Power can then be expressed by the formula

$$1 - \text{cdfn} \left[\frac{\text{fz}}{\text{se(fz)}} - \text{z value}\right] ,$$

where cdfn is the cumulative density function of a normal distribution.

COHEN'S CRITERIA FOR EFFECT SIZE

We placed emphasis on PR(Y, Z) because it determines effect size in accordance with Cohen's Case 1 formulation (Cohen 1988, 412-14). The phrase *effect size* generally means "the degree to which the phenomenon is present in the population," or the degree to which the null hypothesis is false (Cohen 1988, 9-10). We followed the conventional operational definitions of *small*, *medium*, and *large* effect sizes proposed by Cohen to categorize and determine effect size.

If the relative effect of the test treatment improves outcome scores by lowering them (meaning that there is a negative correlation between outcome and treatment variables), the three effect sizes translate to PR(Y, Z) = –.14 for a small effect size, PR(Y, Z) = –.36 for a medium effect size, and PR(Y, Z) = –.51 for a large effect size. Therefore, H_1: PR(Y, Z) = pr = –.14 for a small effect, –.36 for a medium effect, and –.51 for a large effect. The procedure is not restricted to these values; it can be generalized to pr values from –1 to 1.

Rosenthal and Rubin (1982) translated correlation coefficients to binomial effect size displays to arrive at a more intuitive way to understand the magnitude of an experimental effect. For our purposes, pr values of –.14, –.36, and –.51 corresponded respectively, to approximately 12%, 30%, and 44% improvement due to the test treatment.

EVALUATIVE STRATEGY TO COMPARE DESIGNS

An evaluative strategy needed to be implemented to compare the power and sample size estimates of alternative designs that have varying amounts of randomization and cutoff-based assignment. Power will remain unchanged for a given effect size if the value of PR(Y, Z), which we call pr, is held constant for all design types, because pr essentially determines power. One suggested power analytic technique is to take the traditional RCT with equal numbers in both treatment groups as the base design for comparison purposes. This RCT was assumed to incorporate randomization over the entire baseline measure.

There were two major reasons for using the equally balanced RCT as the base design as opposed to any one of the cutoff-based designs. First, because the RCT is more universally known and more intuitively understandable, it is more natural to quantify and classify effect sizes in terms of the conventional RCT. Second, it offers the most stringent comparison, so power and sample sizes for the cutoff designs will be conservative approximations.

Because the RCT with equal numbers in the treatment groups defined and partitioned small, medium, and large effect sizes, it should be remembered that in what follows, pr values equal to –.14 for a small effect, –.36 for a medium effect, and –.51 for a large effect corresponded to the 50/50 RCT only. The other designs have, for a given effect size based on the 50/50 RCT, partial correlations, whose absolute values were lower than those stated for the 50/50 RCT. Hence their power also was lower.

THE MAIN FORMULAS

R. B. Darlington (personal communication, August 17, 1990) showed that

$$PR(Y, Z) = \pm \frac{1}{\sqrt{(1 + H)}},$$

with the sign depending on the relationship between Y and Z (given X), where

$$H = \frac{V(e) \div (B_z)^2}{(P_0P_1)\,(1 - P_0P_1D^2)}$$

in which V(e) signifies the population residual Y variance, B_z signifies the population treatment effect, P_0 and P_1 signify the overall population proportion of cases assigned to control treatment and test treatment, respectively, and D signifies the difference between the population baseline means of the two treatment groups.

The denominator of H changes with the type of design structure, but remains fixed for a given design structure. Specifically, if the overall proportion of test-treated cases is equal across the designs, then only D^2 varies across designs. Therefore only the numerator of H, namely the ratio, $V(e) \div (B_z^2)$, needed to be manipulated to arrive at the desired value for PR(Y, Z) upon which power was based. Because it is this ratio itself that matters and neither V(e) nor B_z alone, without loss of generality, either V(e) or B_z can be assumed fixed to arrive at the desired value of PR(Y, Z) under the alternative hypothesis. For example, if B_z is fixed, then V(e), the residual Y variance when Y is regressed on X and Z, can be algebraically determined to obtain the appropriate value of pr. And pr determined fz which, along with the critical z value, determined power.

If B_z is fixed to equal −1 (even though the treatment effect is almost surely to be some other value), which was what we assumed in the computations, the RCT requires that V(e) = 12.50 for PR(Y, Z) = −.14, V(e) = 1.68 for PR (Y, Z) = −.36, and V(e) = .71 for PR(Y, Z) = −.51. In the cutoff-based designs, both B_z and V(e) took these same values, but pr was lower (in absolute value) as the denominator of H decreased, because D^2 was no longer zero. The program was written in GAUSS (Aptech Systems, Inc. 1988) and is available on request.

CLASSIFYING SMALL, MEDIUM, AND LARGE CUTOFF INTERVALS

A general classification scheme needed to be devised that defined cutoff-based RCTs as having small, medium, and large cutoff intervals to indicate the amount of randomization in a given cutoff-based RCT. The proposed strategy followed here defined the size of the cutoff interval in terms of the increase in power and efficiency of a cutoff-based RCT relative to the nonrandomized, single cutoff-point (RD) design—but without letting too much randomization introduce much loss in ethical and other practical concerns that motivated the use of cutoff-based RCTs in the first place.

A slight degree of (relative) improvement in efficiency and power was represented by a small-sized cutoff interval, a moderate degree of improvement by a medium-sized cutoff interval, and high degree of improvement by a large-sized cutoff interval—without foregoing much ethical or practical loss. Specifically, small, medium, and large cutoff-interval RCTs are defined here as intervals of randomization that include, respectively, 20%, 35%, and 50% of all cases. The power-analytic procedure can be generalized to allow for varying cutoff widths and amounts of randomization.

RESULTS

Tables 1 and 2 show approximately how many subjects are needed in cutoff-based designs. Everything else the same, wider cutoff intervals require fewer subjects to achieve the same level of power. Table 3 shows approximately how many subjects are needed in the conventional RCT. The tables show that the reduction in sample size is much more from a small to medium effect size than from a medium to large effect size. This is because in the Fisher Z method power is a nonlinear function of PR(Y, Z) and hence effect size.

These three tables answer the question, "How many more subjects are needed in a cutoff-based RCT (and a RD design) to reach the same level of power as the conventional RCT?" Consider statistical power of .80. For a small effect, the RD design, the small, medium, and large cutoff-interval RCTs require, respectively, about 2.73, 2.50, and 2.10, and 1.70 times as many cases as the RCT; for a moderate effect, about 2.54, 2.35, and 2.00, and 1.65 as many cases as the RCT; and for a large effect, about 2.34, 2.20, and 1.85, and 1.55 as many cases as the RCT. Sample size ratios are similar for the other levels of statistical power.

Figures 1 through 3 depict the power curves of the five designs for the three types of effect size. The shape of the power functions is virtually identical across effect sizes. The approximate sample sizes required are not linear with respect to power: higher power values tend to warrant proportionately larger total sample sizes.

TABLE 1: Total Sample Sizes Needed for Cutoff-Based RCTs at the .025 Significance Level (one-tailed)

	Proportion of Cases in the Interval of Randomization								
	20% Effect Size			*35% Effect Size*			*50% Effect Size*		
Power	*Small*	*Medium*	*Large*	*Small*	*Medium*	*Large*	*Small*	*Medium*	*Large*
.30	265	40	20	220	33	17	185	28	15
.35	315	47	23	265	40	19	215	33	17
.40	370	54	26	310	47	23	255	38	19
.45	425	62	29	360	53	25	295	44	22
.50	485	70	33	410	60	28	335	49	24
.55	550	79	37	460	66	31	375	55	27
.60	615	88	41	520	75	35	425	61	29
.65	690	98	45	580	83	39	475	68	32
.70	775	110	50	650	93	43	535	76	36
.75	870	123	56	730	104	48	595	85	40
.80	985	138	63	830	118	54	675	96	45
.85	1125	159	71	945	135	61	765	110	50
.90	1315	183	82	1110	155	70	895	127	58
.95	1625	225	100	1360	190	86	1105	155	71

TABLE 2: Total Sample Sizes Needed for the RD Design at the .025 Significant Level (one-tailed)

	Effect Size		
Power	*Small*	*Medium*	*Large*
.30	283	42	21
.35	341	50	24
.40	403	58	28
.45	463	67	32
.50	528	76	35
.55	598	85	40
.60	673	96	44
.65	761	107	49
.70	848	119	54
.75	953	134	60
.80	1078	150	68
.85	1228	171	77
.90	1433	199	89
.95	1753	243	109

TABLE 3: Total Sample Sizes Needed for RCT at the .025 Significance Level (one-tailed)

	Effect Size		
Power	*Small*	*Medium*	*Large*
.30	110	19	11
.35	130	22	12
.40	150	25	13
.45	175	28	15
.50	195	31	16
.55	225	35	18
.60	250	39	20
.65	280	43	22
.70	315	47	24
.75	350	53	26
.80	395	59	29
.85	450	67	32
.90	525	77	37
.95	645	94	44

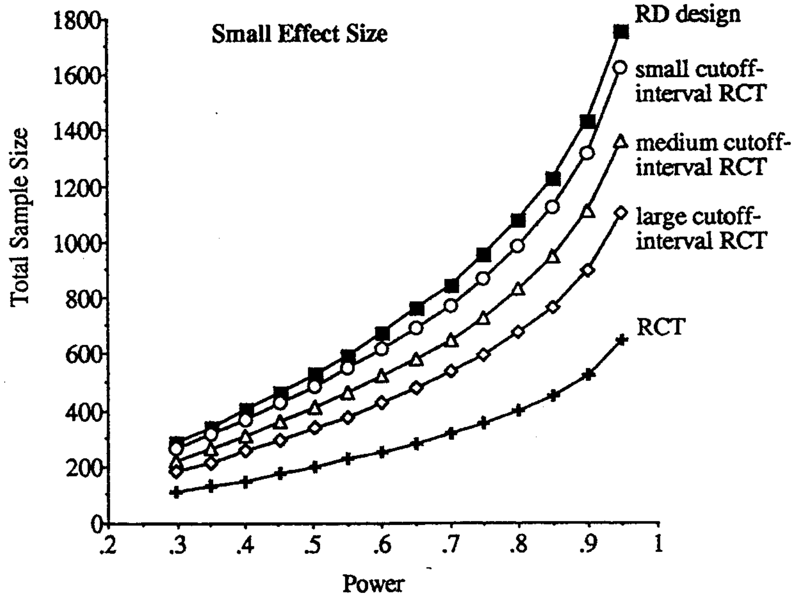

Figure 1: Total Sample Sizes Needed as a Function of Power for a Small Treatment Effect (.025 level of significance, one-tailed test)

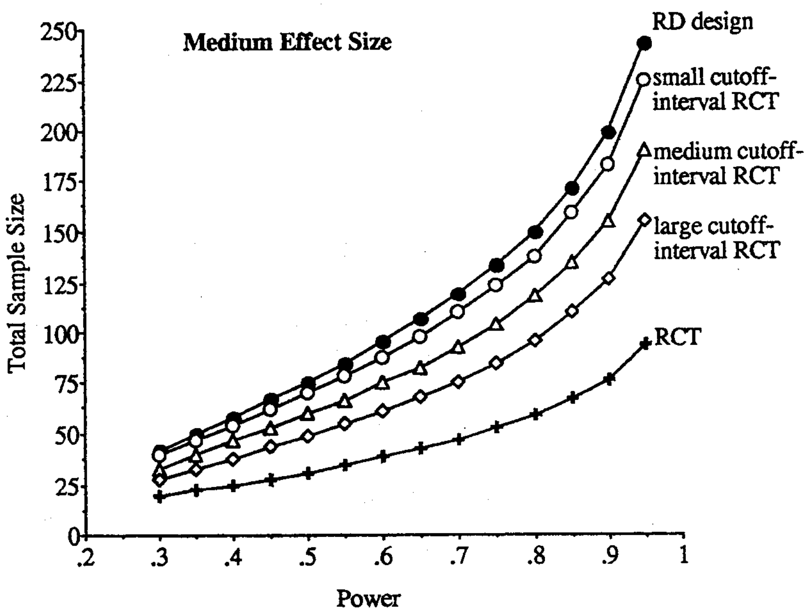

Figure 2: **Total Sample Sizes Needed as a Function of Power for a Moderate Treatment Effect (.025 level of significance, one-tailed test)**

DISCUSSION

Trochim and Cappelleri (1992) discussed that the major benefit of cutoff-based RCTs is the assignment of patients to treatment conditions based on objective, clinically relevant factors. It was also stated that this benefit might be partially offset by the larger samples required for cutoff-based RCTs relative to the conventional RCT. In this article, we provide power tables and curves that will help potential users of cutoff-based RCTs and the conventional RCT, as well as of the RD design, to weigh the tradeoffs that accompany varying levels of randomization and to decide on which design to use.

Sample sizes are extremely sensitive to the expected treatment effect size. Perhaps most important, there is a prominent increase in the total sample size needed—especially for cutoff-based designs—when small treatment effect sizes are expected. Because sample sizes would be drastically different depending on the effect size, it is critical to obtain reliable prior estimates of likely effect size. Clearly, these estimates would be likely to differ across subject areas, settings, types of treatments, and so on. Results of Phase I and II trials may be useful in estimating effect size more accurately.

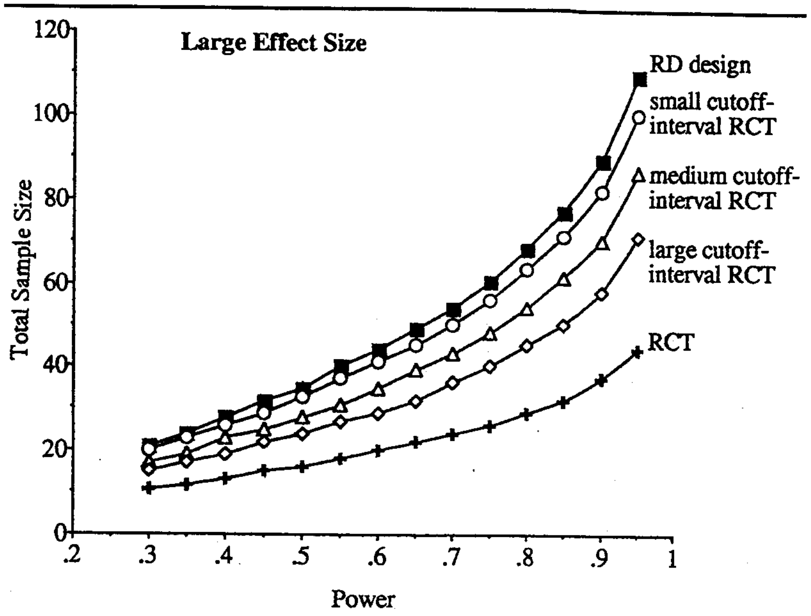

Figure 3: **Total Sample Sizes Needed as a Function of Power for a Large Treatment Effect (.025 level of significance, one-tailed test)**

REFERENCES

Aptech Systems, Inc. 1988. *The GAUSS system version 2.0.* Kent, WA: Aptech Systems, Inc.

Berk, R. A., and D. Rauma. 1983. Capitalizing on nonrandom assignment to treatment: A regression-discontinuity of a crime control program. *Journal of the American Statistical Association* 78:21-28.

Cappelleri, J. C. 1991. Cutoff-based designs in comparison and combination with randomized clinical trials. Ph.D. diss., Cornell University, Ithaca, New York.

Cappelleri, J. C., W.M.K. Trochim, T. D. Stanley, and C. S. Reichardt. 1991. Random measurement error does not bias the treatment effect estimate in the regression-discontinuity design: I. The case of no interaction. *Evaluation Review* 15:395-419.

Cohen, J. 1988. *Statistical power analysis for the behavioral sciences,* 2d ed. Hillsdale, NJ: Lawrence Erlbaum.

Darlington, R. B. 1990. *Regression and linear models.* New York: McGraw-Hill.

Goldberger, A. S. 1972. Selection bias in evaluating treatment effects: Some formal illustrations. Discussion paper #123. Institute for Research on Poverty, Madison, Wisconsin.

Reichardt, C. S., W.M.K. Trochim, and J. C. Cappelleri. Forthcoming. Reports of the death of regression analysis are greatly exaggerated. *Evaluation Review.*

Rosenthal, R., and D. B. Rubin. 1982. A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology* 74:166-69.

Trochim, W.M.K., and J. C. Cappelleri. 1992. Cutoff assignment strategies for enhancing randomized clinical trials. *Controlled Clinical Trials* 13:190-212.

Trochim, W.M.K., J. C. Cappelleri, and C. S. Reichardt. 1991. Random measurement error does not bias the treatment effect estimate in the regression-discontinuity design: II. When an interaction effect is present. *Evaluation Review* 15:571-604.

Joseph C. Cappelleri is a research associate in the Center for Health Services Research and Study Design at New England Medical Center, where he is dealing with methodological and statistical issues in the meta-analysis of randomized controlled trials and evaluating different pharmacological agents in antibiotics, hypertension, congestive heart failure, and AIDS. His interest in research methodology also includes cutoff-based designs, and his substantive interests also include child abuse and neglect. Over the past couple of years, he earned his Ph.D. in evaluation methodology from Cornell University and M.P.H. in quantitative methods from Harvard University.

Richard B. Darlington is a professor of psychology at Cornell University. He has been at Cornell since he received his Ph.D. in 1963 from the University of Minnesota. He is a fellow of the American Association for the Advancement of Science. Professor Darlington has published most extensively on regression and related methods, on the cultural bias of mental tests, and on the long-term effects of preschool programs. He coauthored (with Patricia M. Carlson) the elementary statistics text Behavioral Statistics: Logic and Methods, *published by the Free Press in 1987. His latest book,* Regression and Linear Models, *was published by McGraw-Hill in 1990. He is currently working on a SYSTAT Users Guide.*

William M. K. Trochim is a professor of program evaluation and planning in the Department of Human Service Studies at Cornell University. He has been at Cornell since he received his Ph.D. in 1980 from Northwestern University. He has written widely on quasi-experimental designs and analysis and is the author of the only book-length discussion on the regression-discontinuity design, which was published by Sage in 1984. Professor Trochim has also conducted research on multivariate techniques for conceptualization and pattern matching, and on the use of microsimulation for studying experimental and quasi-experimental designs. In addition, he was awarded by Cornell University for his excellent and distinguished teaching.