

Appendix 2:

# HOW TO CALCULATE SAMPLE SIZE REQUIREMENTS



In Chapter 6, a sample-size table was provided. In this table, the levels of statistical significance and power were set at .90. To compute sample size requirements for different levels of significance and power, the following formula may be used:

$$n = D [Z_{\alpha} (2P (1 - P))^{1/2} + Z_{\beta} (P_1 (1 - P_1) + P_2 (1 - P_2))^{1/2}]^2 / (P_2 - P_1)^2$$

Where:

D = design effect;

$Z_{\alpha}$  = the z-score corresponding to the probability with which it is desired to be able to conclude that an observed change of size  $(P_2 - P_1)$  would not have occurred by chance;

$$P = (P_1 + P_2) / 2;$$

$Z_{\beta}$  = the z-score corresponding to the degree of confidence with which it is desired to be certain of detecting a change of size  $(P_2 - P_1)$ , if one actually occurred.

$P_1$  = the estimated proportion at the time of the first survey; and

$P_2$  = the proportion at some future date such that the quantity  $(P_2 - P_1)$  is the size of the magnitude of change it is desired to be able to detect;

Standard values of  $Z_{\alpha}$  and  $Z_{\beta}$  for use in the above formula are provided in Table 1. A look-up table showing sample sizes needed per survey round for different combinations of significance and power is provided in Table 2.

$\alpha$	$Z_{\alpha}$		$Z_{\beta}$	
	One-Sided Test	Two-Sided Test	$\beta$	$Z_{\beta}$
.90	1.282	1.645	.70	0.53
.95	1.645	1.960	.80	0.84
.975	1.960	2.240	.90	1.282
.99	2.326	2.576	.95	1.645
			.975	1.960
			.99	2.326

**Table 2**  
**Sample Size Requirements for Selected Combinations of  $P_1$ ,  $P_2$ ,  $\alpha$  and  $\beta$**

$P_1$	$P_2$	Combinations of $\alpha$ and $\beta$ ( $\alpha/\beta$ )			
		95/90	95/80	90/90	90/80
.10	.20	432	312	331	228
.10	.25	216	156	165	114
.20	.30	636	460	485	336
.20	.35	299	216	229	158
.30	.40	773	558	594	408
.30	.45	352	255	270	186
.40	.50	841	607	646	444
.40	.55	375	271	288	198
.50	.60	841	607	646	444
.50	.65	367	266	282	194
.60	.70	773	558	594	408
.60	.75	329	238	253	174
.70	.80	636	460	485	336
.70	.85	261	189	200	138
.80	.90	432	312	331	228
.80	.95	163	118	125	86

Note: Sample sizes shown assume a design effect of 2.0.

**What magnitude of change ( $P_2 - P_1$ ) should be measured?**

The quantity ( $P_2 - P_1$ ) is the minimum change in a given indicator that it is desired to measure in successive surveys with a specified degree of certainty. As the value of ( $P_2 - P_1$ ) decreases, the required sample size increases. Thus, for small values of ( $P_2 - P_1$ ), the required sample size will be quite large. Accordingly, for practical reasons, measuring magnitudes of change in behavioral indicators on the order of 10 to 15 percentage points is recommended as the minimum for target group survey efforts, as attempts to measure changes of smaller magnitudes with adequate precision are likely to exceed the resources available in many—in fact, most—efforts.

It will be noted that the magnitude of change parameter specified for sample size determination purposes may or may not correspond to program targets with regard to the indicator in question. In some cases, a program might aspire to change an indicator by only a small amount. For example, where condom use is running only 5 percent in a given setting, it might be quite satisfactory to increase it to 10 percent over a two- to three-year period. Nevertheless, because the sample size required to detect a change of 5 percentage points may be larger than the available resources can support, the parameter ( $P_2 - P_1$ ) might be set to 10 or 15 percentage points in determining sample size requirements for surveys. In this situation, even though the program target of increasing condom use by 5 percentage points may have already been reached, it will not be possible to conclude statistically that the indicator has changed until a change of 10–15 percentage points has been realized (unless, of course, additional resources can be found to support surveys with larger sample sizes).

In cases where larger changes in indicators are expected, it may be desired to increase the magnitude of change parameter in the sample size calculations, thereby decreasing the sample size needed. It should be recognized, however, that doing so will jeopardize the ability to detect smaller changes that may in fact be programmatically significant. For example, if a program aspires to increase condom use by 25 percentage points over a five-year period and accordingly sets ( $P_2 - P_1$ ) equal to 25 percentage points, changes of 10 percentage points realized over the first two years of the program will not be measurable with statistical significance.

Note also that some programs do not have explicit targets for indicators, and thus sample size requirements will be driven primarily by resource and statistical factors. In such cases, the recommended “generic” target of 10–15 percentage points of detectable change is intended as a practical benchmark that should be within the resource levels available for data collection for most programs.

### Determining starting or baseline levels of indicators ( $P_1$ )

A second issue concerns the choice of a starting value of an indicator being monitored, that is,  $P_1$ . Ideally, this choice is based on information available from other surveys conducted in the study setting. Where such information is unavailable, an informed guess must be made. In choosing a value for  $P_1$ , the recommended course of action is to err toward assigning  $P_1$  a value of .50, because the variances of indicators measured as proportions are maximized as they approach .50. Thus, erring toward .50 provides a measure of insurance that the sample size chosen will be sufficient to satisfy the measurement objectives of the survey, even if the estimate of  $P_1$  used is erroneous. The safest course would, of course, be to choose  $P_1 = .50$  for all indicators. However, this would result in samples that are much larger than needed in the event that the actual value of  $P_1$  is very different from .50. Thus, the recommended approach is to make the best guess based on available information, and err toward .50 in selecting values of  $P_1$ .

### Design effects

A third issue concerns the *design effect* ( $D$ ) to be used. The design effect provides a correction for the loss of sampling efficiency, resulting from the use of cluster sampling as opposed to simple random sampling. Thus,  $D$  may be simply interpreted as the factor by which the sample size for a cluster sample would have to be increased in order to produce survey estimates with the same precision as a simple random sample.

The magnitude of  $D$  depends on two factors: (1) the degree of similarity or homogeneity of elements within clusters, and (2) the number of sample elements to be taken from each cluster. The initial factor, the homogeneity of elements within clusters, is a population characteristic over which the survey taker has no control. Prior methodological research indicates that most population characteristics tend to cluster, and thus the prudent course is to assume that some degree of homogeneity within clusters exists. The second parameter, the number of elementary units chosen per cluster, is largely within the control of the survey taker and is an important consideration in the sample design for any survey (see below for further discussion).

What size design effect should be used in estimating sample sizes? Ideally, an estimate of  $D$  for the indicators of interest could be obtained from prior surveys in any given setting. Short of this, “typical” values from surveys conducted elsewhere could be used. If no information is available on the magnitude of design effects for the indicators of interest, the use of a “default” value is recommended. In many cluster surveys, a default value of  $D = 2.0$  is used. Assuming that cluster sample sizes can be kept moderately small in target group survey applications (e.g., not more than 20–25 elements per cluster), the use of a standard value of  $D = 2.0$  should adequately compensate for the use of cluster sampling in most cases.

### **Should one- or two-tailed z-score values be used?**

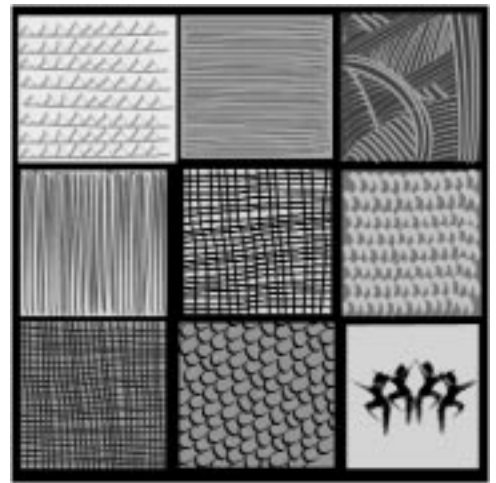
In program evaluation situations, there is good reason to anticipate the direction in which key indicators will change. Accordingly, in the example of sample size computations in Chapter 6, one-tailed values of  $Z_{\alpha}$  were used. This will result in a smaller sample size than if the corresponding two-tailed values had been used. As a general rule, one-tailed tests should be used only when there is a clear rationale for expecting a change in a given indicator in one direction, for example, when an intervention of substantial magnitude and aimed at a given target group has been implemented. Otherwise, the prudent course of action is to use two-tailed values of  $Z_{\alpha}$ .

### **Power**

A point warranting special attention in survey undertakings in which a priority objective is to measure changes in indicators over time is that of *power*. Unless sample sizes are sufficient to be able to detect changes of a specified size, the utility of repeated surveys as a monitoring tool is compromised. To illustrate, suppose we desired to be able to measure a change of 10 percentage points in the proportion of sex workers who always use a condom with their clients. We compare two pairs of hypothetical surveys taken two years apart: one with a sample size of  $n = 500$  per survey round and the other with a sample size of  $n = 200$  per survey round. While both surveys might indicate the expected increase of 10 percentage points, this change may not be statistically significant at a given level of significance based on the survey with a sample size of  $n = 200$ . Thus, we would be forced to conclude that there was no significant change in this behavior over the period study, when in fact there was a real increase that was simply not detectable. To ensure sufficient power, a minimum value of  $Z_{\beta}$  of .80 should be used, and .90 would be preferable if resources permit.

Appendix 3:

# REFERENCE SHELF



Bertrand, J.T., R.J. Magnani and N. Rutenberg. September 1996. *Evaluating Family Planning Programs with Adaptations for Reproductive Health*. Chapel Hill, NC: The EVALUATION Project.

Brindis, C., J.J. Card, S. Niego and J.L. Peterson. 1996. *Assessing Your Community's Needs and Assets: A Collaborative Approach to Adolescent Pregnancy Prevention*. Los Altos, CA: Sociometrics Corporation.

Brindis, C., J.L. Peterson, J.J. Card and M. Eisen. 1996. *Prevention Minimum Evaluation Data Set: A Minimum Data Set for Evaluating Programs Aimed at Preventing Adolescent Pregnancy and STD/HIV/AIDS, 2nd Edition*. Los Altos, CA: Sociometrics Corporation.

Feuerstein, M.T. 1994. *Partners in Evaluation: Evaluating Development and Community Programmes with Participants*. London and Basingstoke: The MacMillan Press Ltd.

Fisher, A., J. Laing, J. Stoeckel and J. Townsend. 1991. *Handbook for Family Planning Operations Research Design, 2nd Edition*. New York: The Population Council.

García-Núñez, J. 1992. *Improving Family Planning Evaluation*. West Hartford, CT: Kumarian Press.

Nelson, K., L. MacLaren and R. Magnani. 1999. *Assessing and Planning for Youth-Friendly Reproductive Health Services*. Washington, DC: FOCUS on Young Adults.

Patton, M.Q. 1990. *Qualitative Evaluation and Research Methods, 2nd Edition*. Newbury Park, CA: Sage Publications.

Rossi, P. and H. Freeman. 1993. *Evaluation: A Systematic Approach, 5th Edition*. Newbury Park, CA: Sage Publications.

Shah, M.K., R. Zambezi and M. Simasiku. 1999. *Listening to Young Voices: Facilitating Participatory Appraisals with Adolescents on Reproductive Health*. Washington, DC: Care International in Zambia and FOCUS on Young Adults.



Appendix 4:

# EVALUATION WEB SITES



## General Sites

**<http://www.eval.org>**

The American Evaluation Association, an international professional association of evaluators, is devoted to the application and exploration of program evaluation, personnel evaluation, evaluation technology and other forms of evaluation.

**<http://www3.sympatico.ca/gpic/gpichome.htm>**

This site offers links to many Web resources on evaluation, brought to you by Government Performance Information Consultants.

**<http://www.unitedway.org/outcomes/>**

The United Way's Resource Network on Outcome Measurement offers a guide to resources for measuring program outcomes for health, human service and youth- and family-serving agencies. Their manual, *Measuring Program Outcomes: A Practical Approach*, can be ordered here.

**<http://www.unites.uqam.ca/ces/mainpage.html>**

The Canadian Evaluation Association is dedicated to the advancement of evaluation for its members and the public. (This site is also available in French.)

**<http://hogg1.lac.utexas.edu/Gen/>**

The Grantmakers Evaluation Network (GEN) is an affinity group of the Council on Foundations. The purpose of GEN is to promote the development and growth of evaluation in philanthropy. GEN will seek to leverage, expand and diversify the sources of philanthropic dollars for evaluation and to build the capacity of members and others in its pursuit.

**<http://www.wmich.edu/evalctr/>**

The Evaluation Center, located at Western Michigan University, is a research and development unit that provides national and international leadership for advancing the theory and practice of evaluation, as applied to education and human services.

**<http://www.socio.com/>**

This is Sociometrics' home page. Click on "Evaluation Resources" for a description of evaluation resources available directly from Sociometrics.

**<http://www.stanford.edu/~davidf/empowermentevaluation.html>**

The American Evaluation Association has a Collaborative, Participatory and Empowerment Evaluation topical interest group that is dedicated to the exploration and refinement of collaborative, participatory and empowerment approaches to evaluation.

**<http://www.inetwork.org/>**

Innovation Network, Inc. (InnoNet), is an organization whose mission is to enable public and nonprofit organizations to better plan, execute and evaluate their structures, operations and services. InnoNet has two services to meet this end: a search service to find model programs, and an evaluation service that guides agencies through a planning and evaluation process. Descriptions of their evaluation methodologies and documents available for ordering are listed on this site.

**<http://trochim.human.cornell.edu/kb/conmap.htm>**

Bill Trochim is a faculty member at Cornell University; his work in applied social research and evaluation is described on this site. His published and unpublished papers, detailed examples of current research projects, useful tools for researchers, an extensive online textbook, a bulletin board for discussions and links to other locations on the Web that deal in applied social research methods are included.

**<http://www.freenet.tlh.fl.us/~polland/qbook.html>**

This site contains a complete manual, entitled *Essentials of Survey Research and Analysis: A Workbook for Community Researchers*, written by Ronald Jay Polland, Ph.D., 1998.

**<http://www.ehr.nsf.gov/EHR/REC/pubs/NSF97-153/start.htm>**

This site contains a complete manual, *User-Friendly Handbook for Mixed Method Evaluations* (August 1997), edited by Joy Frechtling and Laurie Sharp Westat, and developed with support from the National Science Foundation, Division of Research, Evaluation and Communication.

## International Sites

**<http://www.wmich.edu/evalctr/ICCE>**

The International & Cross-Cultural Evaluation Topical Interest Group (I&CCE) is an organization affiliated with the American Evaluation Association. The purpose of the I&CCE is to provide evaluation professionals who are interested in cross-cultural issues with an opportunity to share their experiences with one another.

**<http://www.rrz.uni-koeln.de/ew-fak/Wiso/>**

This is the home page for the German Center of Evaluation (in German) at the University of Cologne. It includes the German translation of the Program Evaluation Standards of the American Evaluation Society.

**[http://www.dec.org/usaaid\\_eval/](http://www.dec.org/usaaid_eval/)**

The U.S. Agency for International Development's Development Experience Clearinghouse (DEC) is a publication clearinghouse that contains references to USAID-funded documentation. The Center for Development Information and Evaluation (CDIE) publications from 1997 through 1998 are provided here and are arranged by CDIE publication series title.

**<http://www.unicef.org/reseval/>**

This site lists some of the monitoring and evaluation tools recently developed by UNICEF and its partners, including the *UNICEF Guide for Monitoring and Evaluation*.

## Education Sites

**<http://ericae.net/>**

This site lists many education-related links for assessment and evaluation.

## Mental Health Sites

**<http://www.vanderbilt.edu/VIPPS/CMHP/>**

The Center for Mental Health Policy is housed in the Vanderbilt Institute for Public Policy Studies of Vanderbilt University, and focuses on child, adolescent and family mental health services research. Their page has links to other mental health-related sites.

## HIV/AIDS Sites

**[http://hivinsite.ucsf.edu/prevention/evaluating\\_programs/](http://hivinsite.ucsf.edu/prevention/evaluating_programs/)**

Maintained by the Center for AIDS Prevention Studies (CAPS) at the University of California, San Francisco (**<http://www.caps.uscf.edu/index.html>**), this site provides tools to help plan, design and implement evaluations of HIV prevention programs.

**<http://www.themeasurementgroup.com/edc.htm>**

The Measurement Group, in collaboration with PROTOTYPES, has been funded by the Health Resources and Services Administration (HRSA) to provide help on evaluation and dissemination activities to 27 national demonstration programs on HIV/AIDS treatment services. This Evaluation and Dissemination Center is part of HRSA's activities to develop innovative models for treating HIV/AIDS.



PHOTO: Harvey Nelson